FACULTY
OF SCIENCES

Ghent University
Biology Department

Royal Belgian Institute of Natural Sciences
Operational Direction Taxonomy and Phylogeny

A GENOTYPE-BY-SEQUENCING (GBS) APPROACH TO UNCOVER GENE FLOW PATTERNS AND HYBRIDISATION SIGNALS IN A SMALL CICHLID GENUS FROM LAKE TANGANYIKA

Jonas Lescroart
Studentnumber: 01307153

Supervisor(s): Dr. Sofie Derycke[1,2]
Prof. Dr. Erik Verheyen[2]
Scientific tutor: Dr. Sofie Derycke

[1] *Biology Department, Ghent University*
[2] *Operational Direction Taxonomy and Phylogeny, Royal Belgian Institute of Natural Sciences*

Master's dissertation submitted to obtain the degree of Master of Science in Biology
Academic year: 2017 - 2018

GHENT
UNIVERSITY

museum

Contents

I. Manuscript

II. Dutch summary

III. Appendix

I. Manuscript

A Genotype-by-Sequencing (GBS) Approach to Uncover Gene Flow Patterns and Hybridisation Signals in a Small Cichlid Genus from Lake Tanganyika

1. Abstract

Genome-wide population surveys are increasingly being employed to gather new insights into the genetic basis of speciation. In this study, we use SNP data generated with genotype-by-sequencing to examine population structure, gene flow patterns and hybridisation signals in *Ophthalmotilapia* spp., a small cichlid genus endemic to Lake Tanganyika. Specimen from the four species were sampled throughout their very different, partially overlapping distribution ranges, allowing detailed inferences about gene flow between both sympatric and allopatric populations, as well as between different species living in sympatry. Furthermore, SNP data enable us to explore how genetic variation is distributed along the genome and which biological processes underlie this variation. We demonstrate that population genomic structure is higher in *O. nasuta* and *O. heterodonta* than in *O. ventralis* and *O. boops*, and make an attempt to link these findings to historic lake level changes in Lake Tanganyika. Whereas our results from intraspecific population genomic analyses comply with conclusions drawn from mitochondrial studies, albeit more detailed, our results from interspecific analyses do not. We find shared genetic variation between each of three *Ophthalmotilapia* species in a region where they live in sympatry, and thus question previous claims based on mitochondrial data that gene flow occurs unidirectional into *O. nasuta*. Instead, our results indicate that gene flow occurs between each of the sympatric species, in both directions. In addition, we calculate pairwise F_{ST}-values for sympatric and allopatric populations of all four species of *Ophthalmotilapia,* and find no clear evidence that would support the hypothesis that genetic variation between diverging populations is organised along the genome in local islands of differentiation. Lastly, we annotate highly divergent SNPs against the latest *Oreochromis niloticus* (Nile tilapia) genome assembly, and find genes that could be involved in biological processes of which the putative role in cichlid speciation was highlighted in previous studies.

2. Introduction

Understanding how populations evolve into distinct species is a central question to the field of evolutionary biology. Multiple mechanisms initiate and enforce diversification, and these are not mutually exclusive. Allopatric divergence (Genner et al. 2007), ecological adaptation (Jones et al. 2012), sexual selection (Seehausen et al. 1997) and hybridisation (Brower 2011, Malinsky et al. 2017) are all recognised as mechanisms that can contribute to the process of speciation. Allopatric divergence refers to the accumulation of genetic differences between populations as a consequence of a physical separation that prevents gene flow (Fitzpatrick et al. 2009). Ecological adaptation arises from the interaction of an organism with other organisms and its environment. Thus, differential adaptational responses can result in ecologically-based divergent selection (Rundle and Nosil 2005). Sexual selection can limit gene flow between populations when it entails assortative mating strategies (Lande 1981). Hybrid genotypes may prove fitter than those of either parent in certain circumstances, or contribute to adaptation through introgression of genetic material (Barton 2001). Of particular interest to the allopatric story of isolation-by-distance is the establishment of secondary contact, and more so when changes in connectivity occur as cyclic events. Climatic oscillations such as during the Pleistocene glaciations may have created the opportunity for populations to go through alternations of allopatric and sympatric distributions (Hewitt 2004). Speciation could presumably be accelerated in this manner, hence the process was described as a 'species pump' (Rossiter 1995, Haffer 2008, Papadopoulou and Knowles 2015).

With the advent of genome-wide population surveys, the idea of a 'genic model of speciation' has amassed great popularity. The model is based on the premises that divergent selection against gene flow is initially restricted to a few genomic regions (Wolf and Ellegren 2017), dubbed 'islands of differentiation' (Via and West 2008). Genes under divergent selection would reduce the effective migration rate of physically linked gene segments, over time resulting in regions of increased local differentiation compared to genome-wide background levels. These regions are referred to as 'islands' and may facilitate speciation-with-gene-flow (Feder et al. 2012), as demonstrated in East African cichlids (Malinsky et al. 2015). Although the idea of differentiation islands is still a matter of debate, and alternative explanations can be given for genomic localised divergence (Wolf and Ellegren 2017), the discussion now focusses on the exact organisation of differentiated regions in genomes of diverging populations. A key issue regarding this debate revolves around the relative importance of different patterns of differentiation observed in genome scans (Feder et al. 2012). Two types are distinguished at opposing ends of the spectrum: islands and continents of differentiation. Islands, as detailed above, denote local peaks of elevated F_{ST} (Via and West 2008), whereas continents stand as metaphor for large-scale inflation of F_{ST}-values, made possible by a global reduction in the effective migration rate of alleles across the genome. This genome-wide reduction could arguably be caused by multifarious divergent selection acting on many loci and affecting many traits throughout the genome (Michel et al. 2010). Feder et al. (2012) propose a four-phased model for speciation-with-gene-flow in which islands and continents are regarded as consecutive phases (phase 2 and 3), preceded by direct selection on individual alleles (phase 1) and culminating in post-speciation divergence (phase 4). The study of hybridisation following secondary contact has provided important insight with respect to this framework (Barton and Hewitt 1989, Vines et al. 2003, Feder et al. 2012).

Lake Tanganyika, situated in the western branch of the East African rift system, is one of the largest and oldest fresh water lakes in the world (Chorowicz 2005). As are the other East African Great Lakes, it is the cradle to a flock of cichlid fish, extraordinary in its number of species, complexity and morphological diversity (Fryer et al. 1972, Sturmbauer 1998). Since its origin 9-12 Mya (Chorowicz 2005), Lake Tanganyika endured a long history of climate changes and complex geological activity, manifested as fluctuating lake levels and dynamic basin morphology (Cohen et al. 1993, Cohen et al. 1997). The present-day lake consists of three subbasins, a feature reflected in its formation from three swampy proto-lakes (Cohen et al. 1997), which deepened over time and fused to a single lake around 5-6 Mya (Cohen et al. 1993). The tropical climate became increasingly drier in the Middle Pleistocene (1.1 Mya) and the lake level dropped by 650-700 m below the current level (Lezzar et al. 1996, Cohen et al. 1997). Following this low stand, the water level of Lake Tanganyika restored and lowered again in alternating cycles, in part coinciding with Late Pleistocene glacial cycles (Cohen et al. 1997). The latest major low stand of at least 435 m below present levels is estimated at 106 Kya (McGlue et al. 2008). A drop of 600 m below the present lake-level should result in isolation of two or three of the lake's subbasins, and the recurrent character of low stand events translated to cycles of cichlid population fragmentation and fusion, which in turn affected population demography and structure (Verheyen et al. 1996, Sturmbauer et al. 2001, Sefc et al. 2007, Nevado et al. 2013, Winkelmann et al. 2017).

The potential for altered population structure and allopatric diversification is most pronounced in stenotopic, shallow water rock dwellers with limited dispersal capability, who exhibit little to no gene flow across habitat barriers such as muddy river mouths and deep-water sections (Sefc et al. 2007, Koblmuller et al. 2011). Moreover, the lake's varied bathymetric shoreline profiles can differentially modulate the effects of lake level fluctuations on cichlid populations living along different stretches of shoreline. Shorelines in the northern and southern parts of Lake Tanganyika drop gradually and have a large shallow area, which is exposed during low stands. In the central part of the lake, the shoreline has a steep inclination, considered a more stable habitat. The former were associated with increased genetic diversity in multiple species of cichlid fish (Koblmuller et al. 2011, Nevado et al. 2013).

The cichlid genus *Ophthalmotilapia* from the tribe Ectodini is endemic to Lake Tanganyika. It lends itself well to investigate effects of historic lake level changes on population structure and speciation dynamics for a number of reasons. First, the different species have different distribution ranges that partially overlap (Hanssens et al. 1999), allowing comparison between both sympatric and allopatric populations and species. Second, the species have similar life histories and ecological needs. They occupy shallow habitats and are restricted by non-rocky stretches acting as habitat barriers (Konings 2014), making them sensitive to lake level changes. Third, hybridisation has been reported within the genus (Kéver et al. 2018) and patterns of introgression have been described based on mtDNA data and nuclear microsatellites (Nevado et al. 2011). Four valid species are currently being recognised: *Ophthalmotilapia nasuta*, which has a circumlacustrine distribution; *O. boops*, which occupies a small stretch of shoreline on the southeastern coast; *O. heterodonta*, distributed along the northern and middle sections of the lake; and *O. ventralis*, restricted to the shorelines of the southern subbasin (Hanssens et al. 1999). Some populations cannot be morphologically assigned with certainty to either of the two latter species, leaving their taxonomy as of yet not fully resolved (Nevado et al. 2011, Konings 2014). In addition, the most northern populations of *O. ventralis* can be discerned by slight morphological differences, and are referred to as *O.* cf. *ventralis* (Konings 2014). Males of all four species of *Ophthalmotilapia* have strongly elongated pelvic fins ending in bifid spatulae, a characteristic unique to this genus within its tribe. Furthermore, males display bright colours when sexually active, in contrast to females, and will guard a breeding ground and perform courtship behaviour to attract females (Hanssens et al. 1999, Kéver et al. 2018). These features are indicative to sexual selection, a prerequisite mechanism for asymmetric behavioural reproductive isolation, which is one of two explanations offered by Nevado et al. (2011) for the unidirectional introgression they observed into *O. nasuta,* the other being cytonuclear incompatibility. Sexual selection was also evoked as explanatory mechanism for female multiple mating in *O. ventralis* (Immler and Taborsky 2009). Although differences in courtship behaviour are documented for *O. nasuta* and *O. ventralis,* these are not likely to constitute a solid prezygotic barrier (Kéver et al. 2018).

Over the past years, sampling efforts have culminated into a comprehensive collection of *Ophthalmotilapia* specimen, available at the Royal Belgian Institute of Natural Sciences and the Royal Museum for Central Africa. For a selection of these, genotype-by-sequencing (GBS) libraries were recently sequenced at the KU Leuven Genomics Core to retrieve genome-wide single nucleotide polymorphism (SNP) data for all four *Ophthalmotilapia* species. SNP data allow for fine-scale characterization of population genomic structure and gene flow patterns (Larson et al. 2014, Alter et al. 2017, Deperi et al. 2018, Peart et al. 2018), and because it is a genome-wide approach, it presents the possibility to examine patterns of differentiation along the genome. In this study, we seek to improve our understanding of the genetic basis of speciation by examining how intraspecific genetic variation translates to genetic variation shared between closely related species living in sympatry. We used SNP data to address the following main questions: 1) what pattern of population genomic structuring is present in all four species of *Ophthalmotilapia* throughout their complete distribution range?; 2) how does this population structure compare to phylogeographic patterns derived from mitochondrial DNA?; 3) how is genetic variation between populations of each species structured across the genome?; and 4) what pattern of hybridisation can be retrieved from genomic data in sympatric species of *Ophthalmotilapia*? In addition, we explore the functionality of the genes containing the SNPs that most prominently differentiate between populations and species. With respect to these questions, we hypothesise that 1) population genomic structures reflect historic lake level changes; 2) population structures derived from genomic data are in agreement with results from mtDNA, but are provided in more detail; 3) genetic variation between populations is organised in regions of increased differentiation; 4) the pattern of hybridisation derived from genomic data is congruent with unidirectional introgression into *O. nasuta*, as was documented based on mtDNA in previous work (Nevado et al. 2011).

3. Materials and Methods

3.1 Sampling

In this study, a total of 616 specimen of *Ophthalmotilapia* spp. were used from 78 different localities around Lake Tanganyika, covering all four species throughout their respective distribution ranges (Table S1). Of these 616 specimen, 402 were made available by the Royal Belgian Institute of Natural Sciences (RBINS) and the Royal Museum for Central Africa (RMCA) for DNA extraction. Tissue samples were kept in ethanol until extraction. Voucher specimen are stored in the RBINS and RMCA. Total DNA was extracted from the samples following standard protocols (NucleoSpin® Tissue, Macherey-Nagel, Düren, Germany).

3.2 *Phylogenetic and phylogeographic analysis using mtDNA*

For 23 *O. boops* specimen, sampled by A. Konings in the vicinity of Nkondwe Island, the first most variable part of the mitochondrial control region was amplified using the PCR protocol as specified in Table S2. PCR products were purified using 0.5 µL Exonuclease I enzyme and 1 µL Thermo Scientific FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher Scientific, Massachusetts, USA) for every 5 µL of PCR mixture. Sequencing reactions were performed in both directions using the forward primer TDK-DH4-short (5'-GATCCCATCTTCAGTGTTATGC-3') and reverse primer LPro$_2$ (5'-CTCTCACCCCTAGCTCCCAAAG-3'), published by Nevado et al. (2009), and run on an ABI 3130 XL sequencer following the manufacturer’s protocol. Forward and reverse sequences were assembled and checked by eye using the *SeqMan* program included in the *DNASTAR* software *v7.1.0* (DNASTAR Inc., Madison, WI, USA). Only the reverse sequence was used for sample ‘OB_02012018_15’ due to a poor quality forward read. In addition, 467 partial D-loop sequences were obtained from previous studies conducted at the RBINS and RMCA involving *Ophthalmotilapia*, yielding a mtDNA dataset of

490 partial D-loop sequences from 74 localities across Lake Tanganyika. This dataset contained 77 *O. boops*, 226 *O. nasuta*, 62 *O. heterodonta* and 125 *O. ventralis* sequences of which 13 were labelled as *O.* cf. *ventralis*. Alignment was performed with the *MUSCLE* feature in *MEGA v7.0.20* (Kumar et al. 2016) . The aligned sequences were trimmed to fragments of 307 bp. In order to assess the population structure of *Ophthalmotilapia* species, haplotype networks with all 490 partial D-loop sequences were constructed in *PopART v1.7* (Leigh and Bryant 2015) using median-joining inference (Bandelt et al. 1999). Gaps were masked and sequences that contained significantly more gaps than others were kept, not discarded. Haplotype networks were coloured either by taxonomic affinity or by geographic region. The geographic regions are based on the three large subbasins present in Lake Tanganyika, because of the likely historical importance of the subbasins in shaping connectivity between populations of *Ophthalmotilapia* species. These regions are delineated as the eastern sides of the southern subbasin (ES), the central subbasin (EC) and the northern subbasin (EN) and as the western sides of the southern subbasin (WS), the central subbasin (WC) and the northern subbasin (WN). In Figure 1, the regions correspond to localities 1-18 (ES), 19-23 (EC), 24-27 (EN), 30-38 (WS), 40-42 including 53 (WC) and 43-51 (WN). The six regions are outlined in Figure 2 (see Results). Grouped per region, the D-loop dataset contains 223 ES, 45 EC, 31 EN, 82 WS, 9 WC and 100 WN sequences.

3.3 Population genomic analysis using genotyping-by-sequencing data

For the genomic dataset, 384 individuals from 69 localities were used, sampled across Lake Tanganyika, comprising all *Ophthalmotilapia* species (69 *O. boops*, 54 *O. heterodonta*, 141 *O. nasuta* and 115 *O. ventralis*) and three outgroup species (one *Callochromis pleurospilus*, two *Ectodus descampi* and two *Xenotilapia ochrogenys*) from the Ectodini tribe. Localities were generally pooled within 0.1 decimal degree, although exceptions were made where needed to obtain a minimum of five individuals per pooled locality. This resulted in 46 pooled localities (Figure 1, Table S1). DNA quality was assessed using agarose gel electrophoresis. Using the genotype-by-sequencing (GBS) protocol as described in Elshire et al. (2011), two libraries were prepared with each 192 barcoded samples in the Genomics Core facility (KULeuven, Belgium). In short, the samples were digested with PstI and the digested samples with sticky ends were ligated to barcoded adaptors. Following this procedure, the samples were cleaned with AMPure beads, amplified with PCR and cleaned again. The amount of template present in the samples for sequencing was quantified by measuring with the PicoGreen. Samples were equimolarly pooled with 192 samples per library. The two genomic libraries were sequenced on two lanes of an Illumina HiSeq (125 bp, PE).

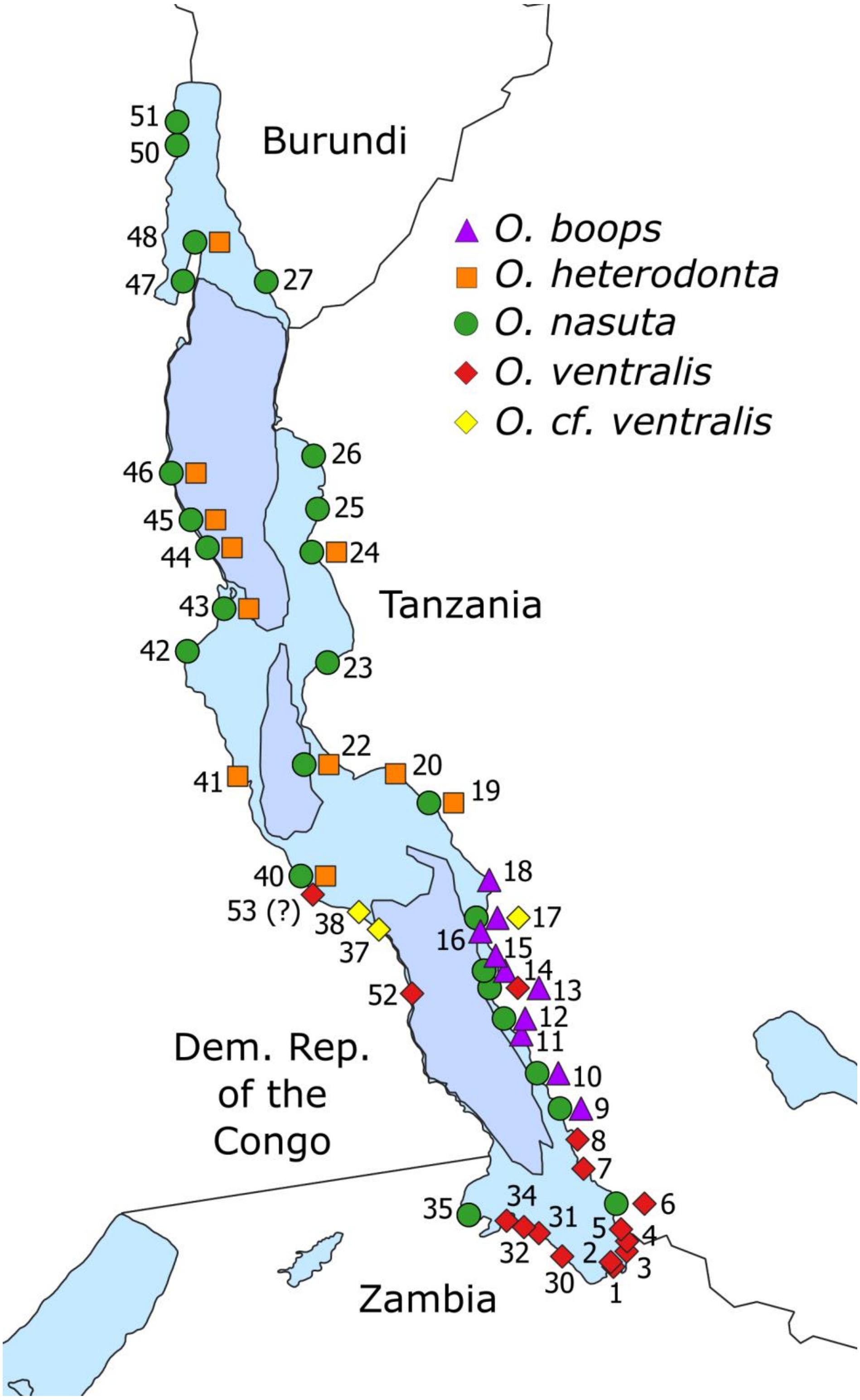

Figure 1. Map of Lake Tanganyika with the 46 pooled localities sampled for the genomic dataset in this study. Apart from locality 53, pooled localities represent at least five individuals. Locality 53 is flagged with a question mark because it is associated with only one specimen and that specimen was obtained through aquarium trade. Therefore the accuracy of the specimen's origin cannot be guaranteed. The numbering of localities is not necessary continuous. Taxonomic sampling diversity is given for each locality, and roughly indicates the species' distribution ranges. Outgroup species are not displayed for clarity. Subbasins are shown in greyish blue.

3.3.1 Filtering and quality control

Demultiplexing was done with the *GBSX* software (Herten et al. 2015) and overlapping reads were merged to fragments with the *FLASH* software (Magoc and Salzberg 2011). *Bowtie 2* (Langmead and Salzberg 2012) was used to align the reads to an *Oreochromis niloticus* genome (assembly ASM185804v2) which was downloaded from *GenBank* (Agarwala et al. 2018). The *GenBank* assembly accession number is GCA_001858045.2. Variants were called using the *FreeBayes* software (Garrison and Marth 2012), with cut-off values set to 15 for both mapping quality and base quality. SNP data were exported in VCF format (Danecek et al. 2011) for downstream analyses. Above part of the bioinformatics work flow was conducted by the KU Leuven Genomics Core. The resulting dataset was filtered using *VCFtools v.0.1.14* (Danecek et al. 2011). Since the order in which filtering steps are applied in *VCFtools* matters, the order as described below was respected for all analyses. First, the five outgroup individuals were removed. These were included in the sequencing runs for analyses not presented in this study. Next, individuals that had < 10% of SNPs genotyped were discarded, as well as SNPs that were genotyped in < 50% of individuals. Further, the dataset was filtered to retain only biallelic SNP sites, as this was required for downstream analyses. From the remaining SNPs, all sites deemed homozygote alternatives (with one non-reference allele covering > 99% of sites of all individuals) were excluded. Sites constituting an indel with respect to the reference genome were removed as well. Finally, SNPs were filtered based on read depth and genotype quality. Different settings for depth filtering were explored and the threshold to retain an individual site was set to a minimum depth of 20 for each site. This level of stringency was applicable for lake-wide exploratory analyses drawing from all available individuals, however it resulted in insufficiently large SNP datasets when we tried to work with subsets of individuals (e.g. a single species). To take advantage of larger SNP datasets when working with subsets, and in order to capture more variation needed for finer-scale analyses, the reading depth threshold to retain sites was set to a minimum average depth of 10 for that site over all included individuals. For the complete dataset and all subsets, genotypes with a quality score < 30 were excluded. Five subsets were made and filtered separately with *VCFtools*: one monospecific subset for each of the four *Ophthalmotilapia* species, and one subset containing all individuals sampled on the eastern side of the southernmost subbasin of Lake Tanganyika (ES), a region where *O. boops*, *O. nasuta* and *O. ventralis* (including *O.* cf. *ventralis*) live in sympatry (Table S1).

3.3.2 *Intraspecific population genomic analysis of the four* Ophthalmotilapia *species*

To investigate the population genomic structure for each species of *Ophthalmotilapia* throughout their respective distribution ranges, we performed a discriminant analysis of principal components (DAPC) for each species. DAPC is a multivariate statistical approach that clusters genetically related individuals whilst maximising discrimination between groups (Jombart et al. 2010). DAPCs for the four monospecific subsets were performed in *R v.3.4.2* (R Development Core Team 2008), using the *R* packages *'vcfR' v1.8.0* (Knaus and Grunwald 2017) and *'adegenet' v2.1.1* (Jombart 2008) along with a number of aesthetical packages (*'ggplot2', 'reshape2', 'igraph', 'RColourBrewer'*). *R* scripts were modified from a publically available tutorial (Jombart 2015). The optimal number of PCs to retain for a given DAPC was determined by computing a-scores (Jombart 2008). Following convention, the number of retained discriminant eigenvalues for every DAPC was set to the number of clusters minus one. Based on the DAPCs, two outliers (shortnames 'heterodonta_23_AATGT' and 'ventralis_14_GAATCTCA') were excluded from these and further analyses because they fell completely outside the vector space occupied by conspecific samples. These two samples were possibly misidentified, but for neither of them a voucher specimen is available to confirm identification. One additional individual (shortname 'ventralis_36_TAGCACA') was removed from the *O. ventralis* subset because its sampling coordinates were uncertain. To quantify gene flow between conspecific populations across the lake, individual posterior membership probabilities as computed in the DAPCs were plotted in a compoplot for each species, using the same *R* packages.

A more in-depth look at the biological processes responsible for population genetic structuring as it is observed within *Ophthalmotilapia* species was attained by identifying and annotating the SNPs that contributed the most to the DAPC clustering of genotypes. This analysis was carried out on the *O. nasuta* subset, because it is the only *Ophthalmotilapia* species with a lake-wide distribution. This implies a stronger isolation-by-distance effect, which in turn leads us to expect a higher degree of population structure in *O. nasuta* compared to its congenerics. A DAPC was already performed for this subset (cf. supra), and from that analysis, loading values can be extracted for all SNPs in that subset. These values indicate the relative contribution a SNP has to the genetic clustering and discrimination along a given linear discriminant of the DAPC. The SNPs with the highest loading values along the first two linear discriminants were identified in *R* using an arbitrarily chosen threshold of > 0.0012, with the aim of retaining between 50 and 100 SNPs in total. Again, *R* scripts were modified from the previously mentioned publically available tutorial (Jombart 2015) and require the same packages. These high-profile SNPs were then annotated with *SNPdat v1.0.5* (Doran and Creevey 2013) with the *Oreochromis niloticus* genome (assembly ASM185804v2) as reference genome and annotation followed the NCBI *O. niloticus* Annotation Release 103. Both the reference genome (assembly accession number

GCA_001858045.2) and its annotation were retrieved from *GenBank* (Agarwala et al. 2018). The *SNPdat* annotation of genes containing these SNPs was supplemented with *Ensembl* Gene Stable IDs and Gene Ontology (GO) term names (Harris et al. 2004), both retrieved from *Ensembl* (Hubbard et al. 2002) through their *BioMart* feature, to provide an external gene identifier and to provide a notion of the biological function of the genes. Only GO term names that classified under the GO domain 'biological process' were retained, thereby discarding any GO term name classified as 'cellular component' or 'molecular function'. Note that *Ensembl* makes use of a different *Oreochromis niloticus* assembly (assembly Orenil1.1, *GenBank* accession number GCA_001858045.2) than the assembly used by *SNPdat*.

3.3.3 Genome-wide differentiation patterns within the four Ophthalmotilapia *species*

In order to examine whether differences exist in the amount of differentiation that SNPs display along the genome, Weir and Cockerham F_{ST}-values (Weir and Cockerham 1984) were calculated using *VCFtools* (Danecek et al. 2011) and visualised with manhattan plots using the *'qqman'* package *v0.1.4* (Turner 2014) in *R*. Negative F_{ST} -values were set to zero. This was done for pairs of intraspecific populations throughout Lake Tanganyika. We selected a pair of populations that live in sympatry and a pair of populations that live in allopatry for each *Ophthalmotilapia* species (Table 1). Populations were selected with the aim of minimising and, respectively, maximising geographic and genetic distance between paired populations based on sampling locality and DAPC outcome. Sample size per population was also taken into consideration. Since *O. boops* has a small distribution range, all subpopulations could be considered sympatric, and for the pair that we dubbed 'allopatric', we resorted to simply selecting the localities that are situated at opposite ends of the species' distribution range.

Species	**Distribution**	**1st locality and (number of ind.)**	**2nd locality and (number of ind.)**	**Number of SNPs with F_{ST} > 0**
O. boops	sympatric	11 (9)	12 (7)	1104
	'allopatric'	10 (8)	17 (13)	486
O. nasuta	sympatric	44 (10)	45 (10)	1764
	allopatric	6 (15)	48 (9)	1263
O. heterodonta	sympatric	43 (8)	45 (8)	1263
	allopatric	19 (8)	45 (8)	555
O. ventralis	sympatric	31 (5)	32 (7)	2483
	allopatric	6 (24)	52 (6)	645

Table 1. Localities selected for pairwise Weir and Cockerham F_{ST} calculations. Number of individuals is shown in brackets for each locality. Number of SNPs that yielded an FST-value greater than zero is shown in the rightmost column.

3.3.4 Exploration of hybridisation signals between three sympatric Ophthalmotilapia *species*

Hybridisation has been reported between *Ophthalmotilapia* species (Nevado et al. 2011, Kéver et al. 2018), but this research was based on mtDNA and a limited number of genomic markers. To determine whether the same signals of hybridisation that these approaches unveiled can also be uncovered with SNP data on a genome-wide scale, the subset containing the individuals from the sympatric region was analyzed. An interspecific DAPC was carried out for this subset. Posterior membership probabilities were calculated and plotted in a compoplot to examine gene flow between sympatric species. These analyses follow the methodology as outlined previously. Lastly, we wanted to explore the possible biological processes that drive differentiation between sympatric *Ophthalmotilapia* species. To this end, SNPs with high loading values were identified and annotated for two groups of individuals. One group was the sympatric subset. The other group was a version of that subset where all individuals that do not cluster according to their morphological species assignment were removed, i.e. all individuals that had a higher posterior membership probability for a species other than the one they were assigned to morphologically, were excluded. Methodology for SNP identification and annotation follows the procedure as outlined above, except that SNPs were retained based on a loading value threshold of > 0.005 instead of > 0.0012.

4. Results

4.1 *Phylogenetic and phylogeographic analysis using mtDNA*

The 490 D-loop sequences of the mtDNA dataset consist of 131 unique haplotypes when taking nucleotide insertions and deletions into account. Gaps were masked however, resulting in a collapse to 70 haplotypes (Figure 2). In calculating a median-joining network, *PopART* (Leigh and Bryant 2015) identified 40 segregating sites, of which 28 were parsimony-informative sites. Haplotypes were separated from neighbouring haplotypes by one to maximum three mutations. In the haplotype network coloured by taxonomic affinity, roughly three clusters can be distinguished. Sequences of individuals from the same species grouped mostly together, with *O. boops* and *O. nasuta* sequences both forming rather distinct clusters, while *O. ventralis* (including *O.* cf. *ventralis*) and *O. heterodonta* sequences make up a third cluster. A number of *O. nasuta* and even a few *O. ventralis* sequences are associated with the *O. boops* cluster. The *O. nasuta* cluster is fairly well defined, in contrast to the *O. ventralis – O. heterodonta* complex, which shares haplotypes with both *O. boops* and *O. nasuta* individuals. In the haplotype network coloured by geographic region, individuals from the previously mentioned *O. boops* cluster all originate from the eastern side of the southern subbasin (ES). Sequences grouped in the *O. nasuta* cluster seem only slightly partitioned by their geographic region of origin. The *O. ventralis* individuals belonging to the *O. ventralis* – *O. heterodonta* cluster are a mix of individuals from both sides of the southern subbasin (ES and WS). The *O. heterodonta* individuals from the central regions (EC and WC) are fewer mutational steps apart from the *O. ventralis* individuals than those of the northern regions (EN and WN).

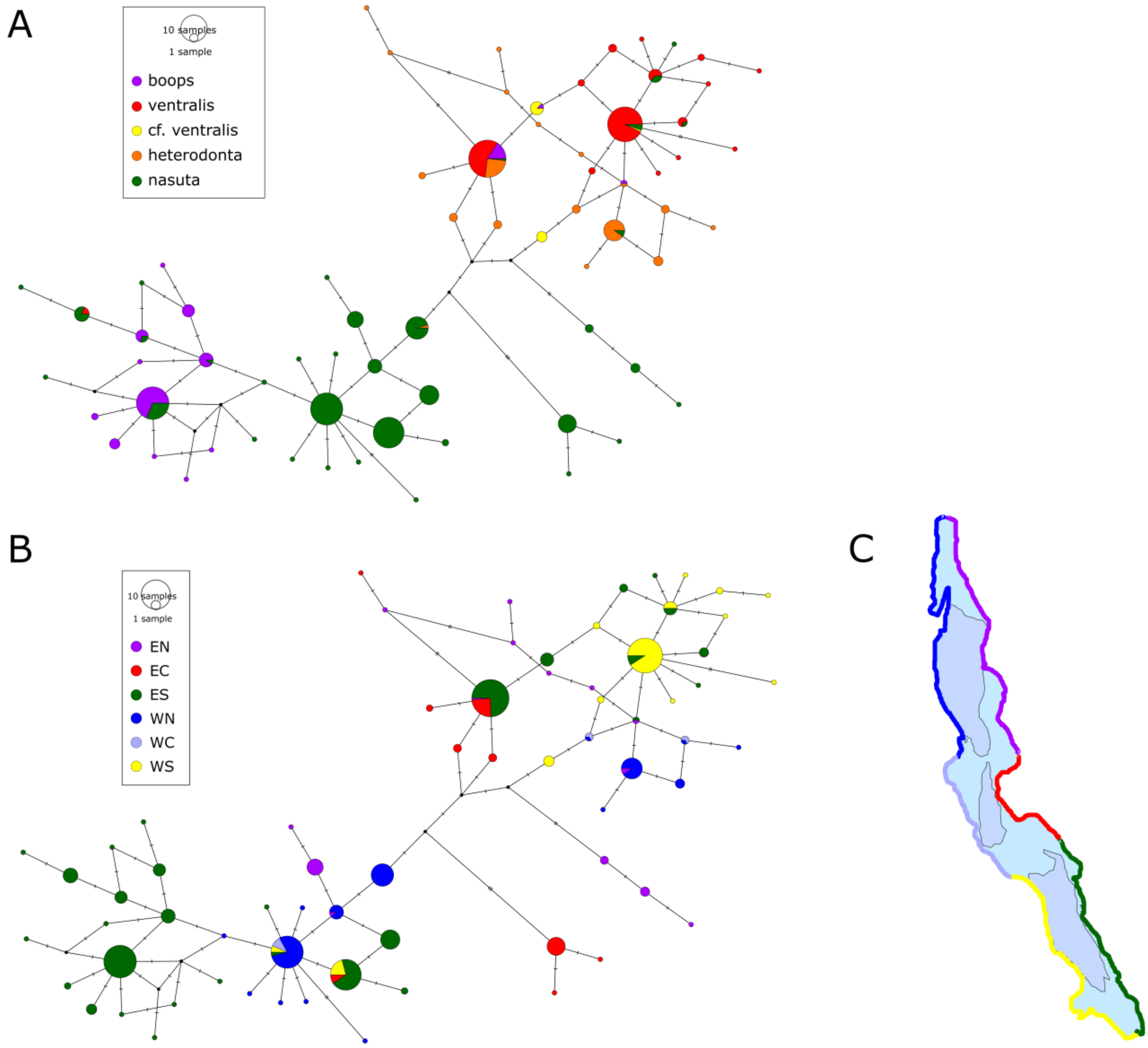

Figure 2. Median-joining haplotype network of 490 mitochondrial D-loop sequences, sampled from four species of *Ophthalmotilapia*. Node sizes correspond to the amount of sequences that were collapsed to the haplotype of a given node. Black nodes represent inferred ancestral haplotypes. A) Haplotype network coloured by taxonomic affinity. Colour codes are given in the upper left corner. B) The same haplotype network as in (A), but coloured by geographic region. Colour codes are given in the upper left corner. C) Map of Lake Tanganyika, indicating the six geographic regions. Colours correspond to (B). Subbasins within the lake are shown in greyish blue.

4.2 Population genomic analysis using genotyping-by-sequencing data

4.2.1 Filtering and quality control

Once sequencing, mapping and variant calling was done, the SNP dataset contained 1 038 666 SNPs. Filtering for missing data caused the largest drop in the number of SNPs (Figure 3). Filtering for reading depth also sharply decreased the number of SNPs, to as few as 357 SNPs, depending on the exact setting. Based on the results in Figure 3, we chose to filter the five subsets used in further analyses for a minimum mean reading depth of 10 for each site over all individuals (meanDP10). After filtering with *VCFtools*, the monospecific subsets retained 68 individuals from 10 sampling localities and 33 411 SNPs (*O. boops*), 140 individuals from 25 sampling localities and 53 400 SNPs (*O. nasuta*), 52 individuals from 11 sampling localities and 79 266 SNPs (*O. heterodonta*) and 112 individuals from 19 sampling localities and 24 703 SNPs (*O. ventralis including O.* cf. *ventralis*). The sympatric region (ES) subset retained 198 individuals from 18 sampling localities and 31 914 SNPs. In addition, two outlier individuals and an individual of unknown sampling locality were removed during further analysis (see Materials and Methods), bringing the final count for the *O. heterodonta*, *O. ventralis* and sympatric subsets to respectively 51, 110 and 197 individuals. Removing these specimen did not affect the number of SNPs.

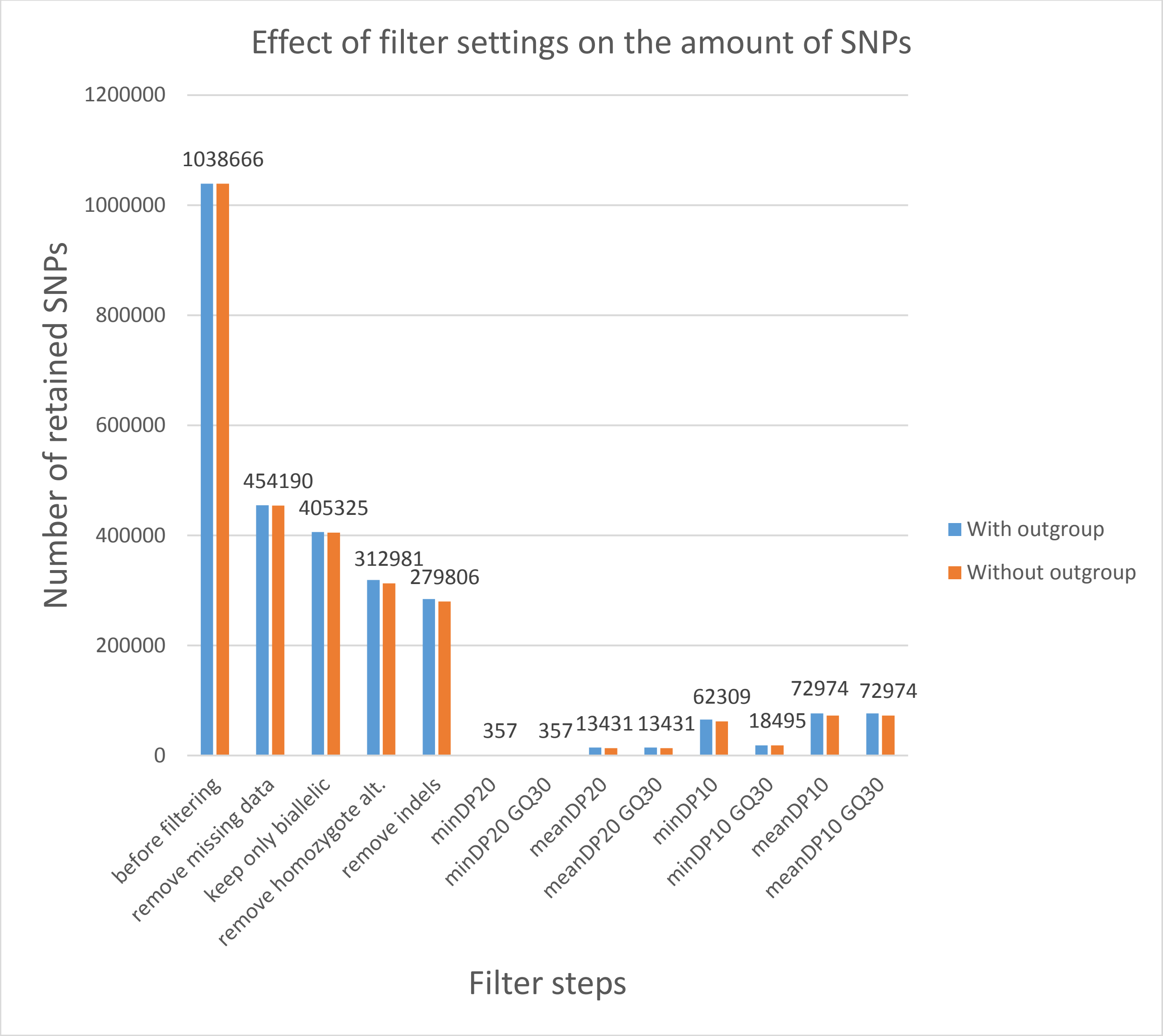

Figure 3. A bar plot displaying the effect of varying filter settings in *VCFtools* on the amount of SNPs that are retained with each step, for the complete dataset. The vertical axis represents the number of SNPs. Filter steps are specified along the horizontal axis. These are ordered chronologically up to the 'remove indels' step, beyond which four different settings for depth filtering (DP) are explored that all follow on the 'remove indels' step. These settings are: minimum depth 20 for all sites (minDP20), minimum mean depth 20 for a given site over all individuals (meanDP20), minimum depth 10 for all sites (minDP10) and minimum mean depth 10 for a given site over all individuals (meanDP10). Each of these settings is followed by a step that filters for genome quality (GQ30). Blue bars represent the number of retained SNPs when the outgroup species were included in the dataset, orange when the outgroup species were excluded. The exact number of SNPs for the dataset without outgroup is given atop of each orange bar.

4.2.2 *Intraspecific population genomic analysis of the four* Ophthalmotilapia *species*

Based on a-scores (Figure S1), the first 8 PCs were retained for *O. boops*, the first 7 PCs for *O. nasuta*, the first 10 PCs for *O. heterodonta* and the first 7 PCs for *O. ventralis*, conserving respectively 33.0%, 44.7%, 69.0% and 24.8% of variance. In *O. boops* three clusters are identified (Figure 4A). Linear discriminant 1 (LD 1) separated among localities 9-14 and 15-18. LD 2 mainly isolates locality 14, plotted below the other clusters. Genotypes within localities are spread out over distances comparable to the distances between clusters. In *O. nasuta*, six clusters can be distinguished (Figure 5A). Locality 27 is separated from all other localities along both LD1 and LD2. 16 out of 25 localities are densely clustered together in the center of the DAPC. Localities 47-48 and 50-51 cluster together and are situated to the left, while 22-23 form a cluster on the right of LD 1. Finally, localities 25 and 26 are separated and spread out to the right of LD 1. The same observations can be made along LD 2. Within localities, variation is restricted. Sampling localities of *O. heterodonta* (Figure 6A) are mostly separated, and scattered to the left of localities 43-45, the only coherent cluster. As with *O. nasuta,* variation within localities is small. In *O. ventralis,* arguably four clusters can be discerned (Figure 7A). Locality 1 and most localities with the numbers 13 and higher, cluster somewhat in the center of both LD 1 and LD 2, with localities 2-8 and 53 gravitating around this center. Along LD 1, localities 7-8 and 53 are spread out to the left of this central cluster, and 2-4 are spread to the right. Along LD 2, localities 5-6 are situated below the central cluster, and 2 and 53 are clearly positioned above this cluster. Genotypes within most localities display considerable variation.

Posterior membership probabilities for the *O. boops* subset (Figure 4B) indicate high admixture, in particular throughout localities 9-13 and within localities 15-18. The posterior membership probabilities for *O. nasuta* (Figure 5B) show that localities 6 through 17 are completely admixed. Individuals from localities 22 to 27 are more sharply segregated, 42 to 46 show progressive admixture, and 47-48 and 50-51 both form delineated pairs. Most samples of *O. heterodonta* (Figure 6B) show very little admixture compared to their congenerics. For *O. ventralis* (Figure 7B), localities 2-8 show distinctly less admixture than the progressive admixture that is observed from locality 13 up to 38.

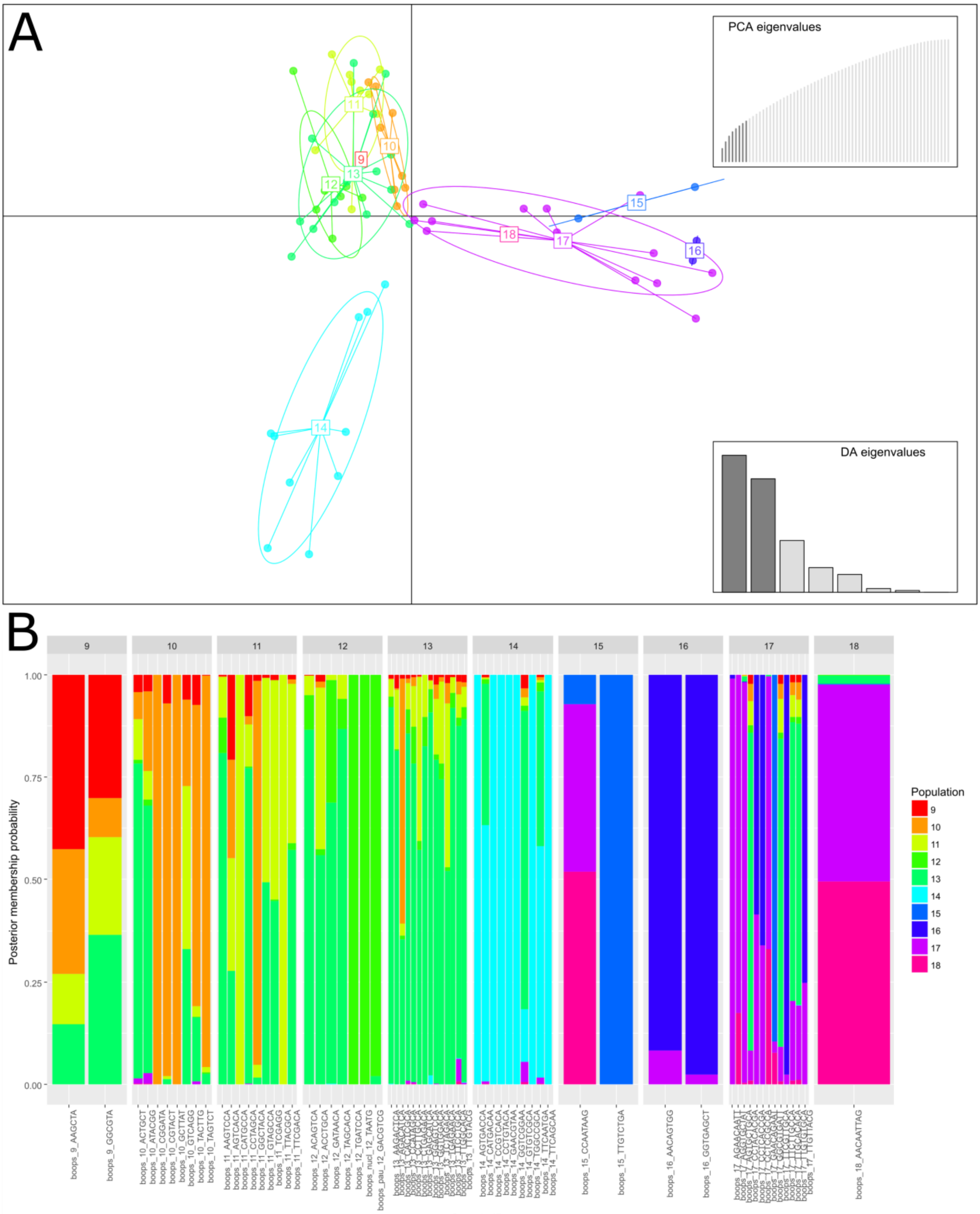

Figure 4. DAPC and resulting posterior membership probabilities for 68 specimen of *O. boops* sampled from 10 pooled localities, based on 33 411 SNPs. A) DAPC where the horizontal axis represents the first linear discriminant, the vertical axis represents the second linear discriminant. Individuals (dots) are coloured and clustered based on the locality they were sampled. Localities are specified by their number. B) Compoplot displaying the posterior membership probabilities of each *O. boops* individual as a bar, cumulatively stacked from 0 to 1. Individual sample IDs for each bar and number identifiers for

each locality are provided below and above the compoplot. Colour codes for the localities are shown on the right and correspond to the colours used in the DAPCs.

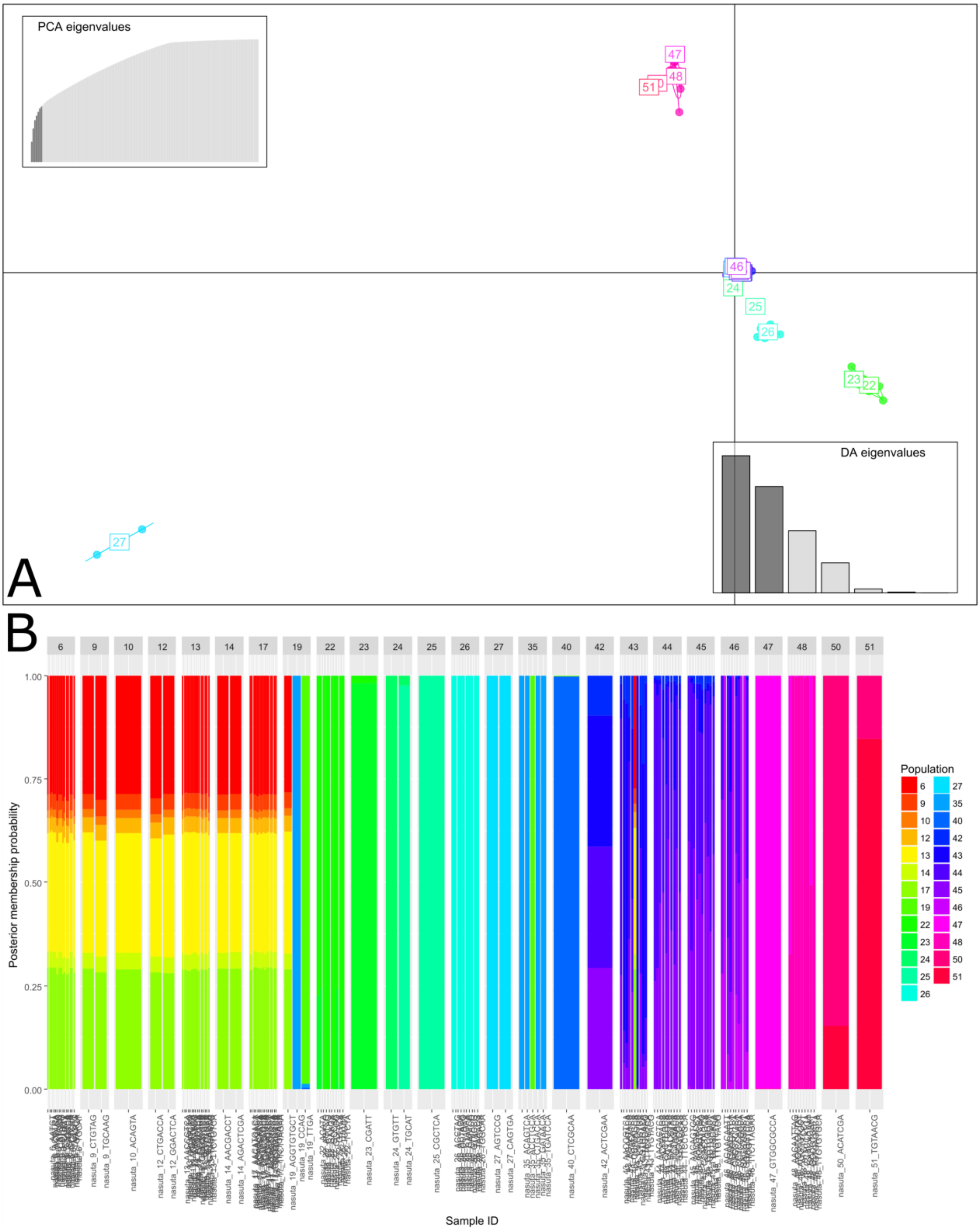

Figure 5. DAPC and resulting posterior membership probabilities for 140 specimen of *O. nasuta* sampled from 25 pooled localities, based on 53 400 SNPs. A) DAPC where the horizontal axis represents the first linear discriminant, the vertical axis represents the second linear discriminant. Individuals (dots) are coloured and clustered based on the locality they were sampled. Localities are specified by their number. B) Compoplot displaying the posterior membership probabilities of each

O. nasuta individual as a bar, cumulatively stacked from 0 to 1. Individual sample IDs for each bar and number identifiers for each locality are provided below and above the compoplot. Colour codes for the localities are shown on the right and correspond to the colours used in the DAPCs.

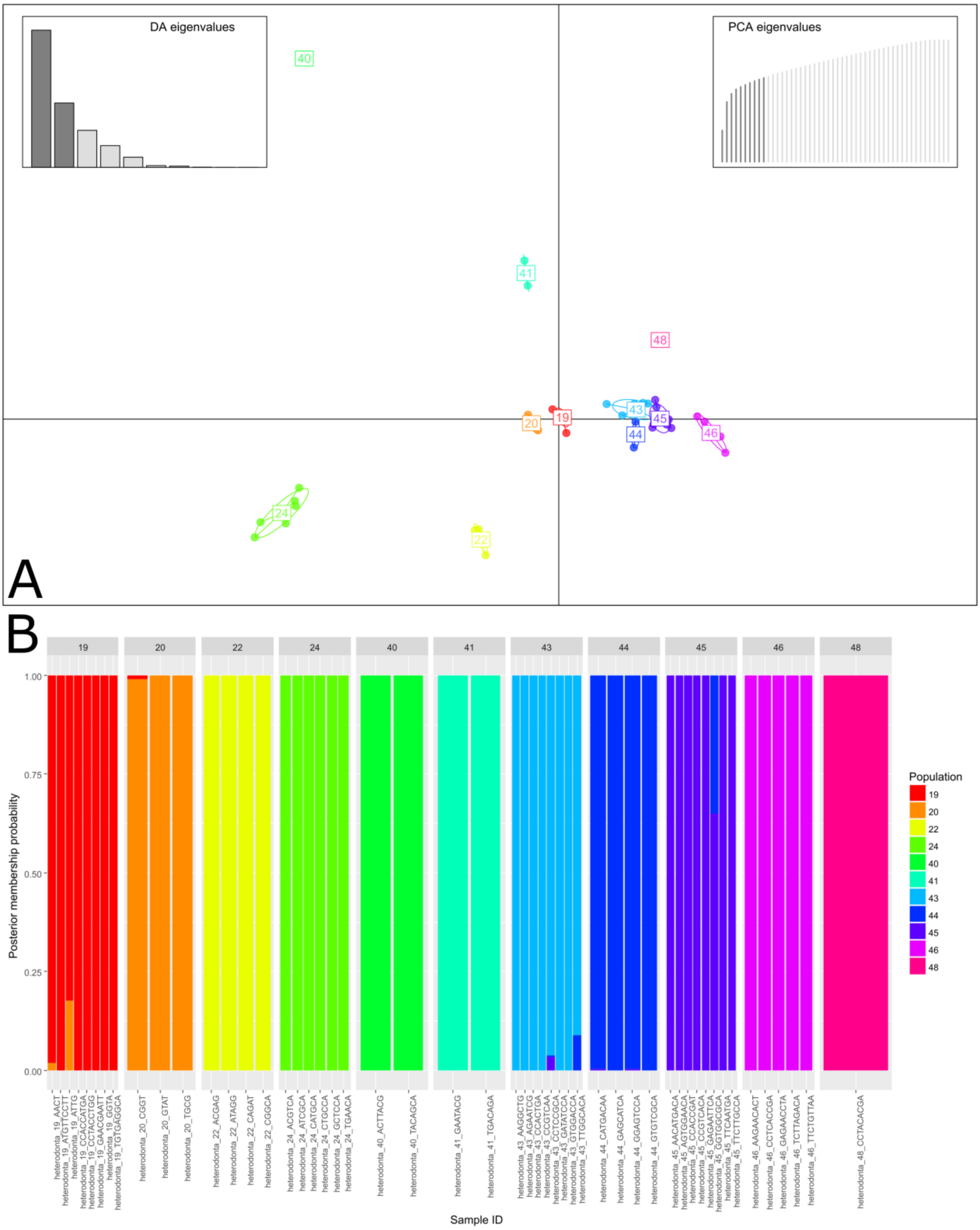

Figure 6. DAPC and resulting posterior membership probabilities for 52 specimen of *O. heterodonta* sampled from 11 pooled localities, based on 79 266 SNPs. A) DAPC where the horizontal axis represents the first linear discriminant, the vertical axis represents the second linear discriminant. Individuals (dots) are coloured and clustered based on the locality they were

sampled. Localities are specified by their number. B) Compoplot displaying the posterior membership probabilities of each *O. heterodonta* individual as a bar, cumulatively stacked from 0 to 1. Individual sample IDs for each bar and number identifiers for each locality are provided below and above the compoplot. Colour codes for the localities are shown on the right and correspond to the colours used in the DAPCs.

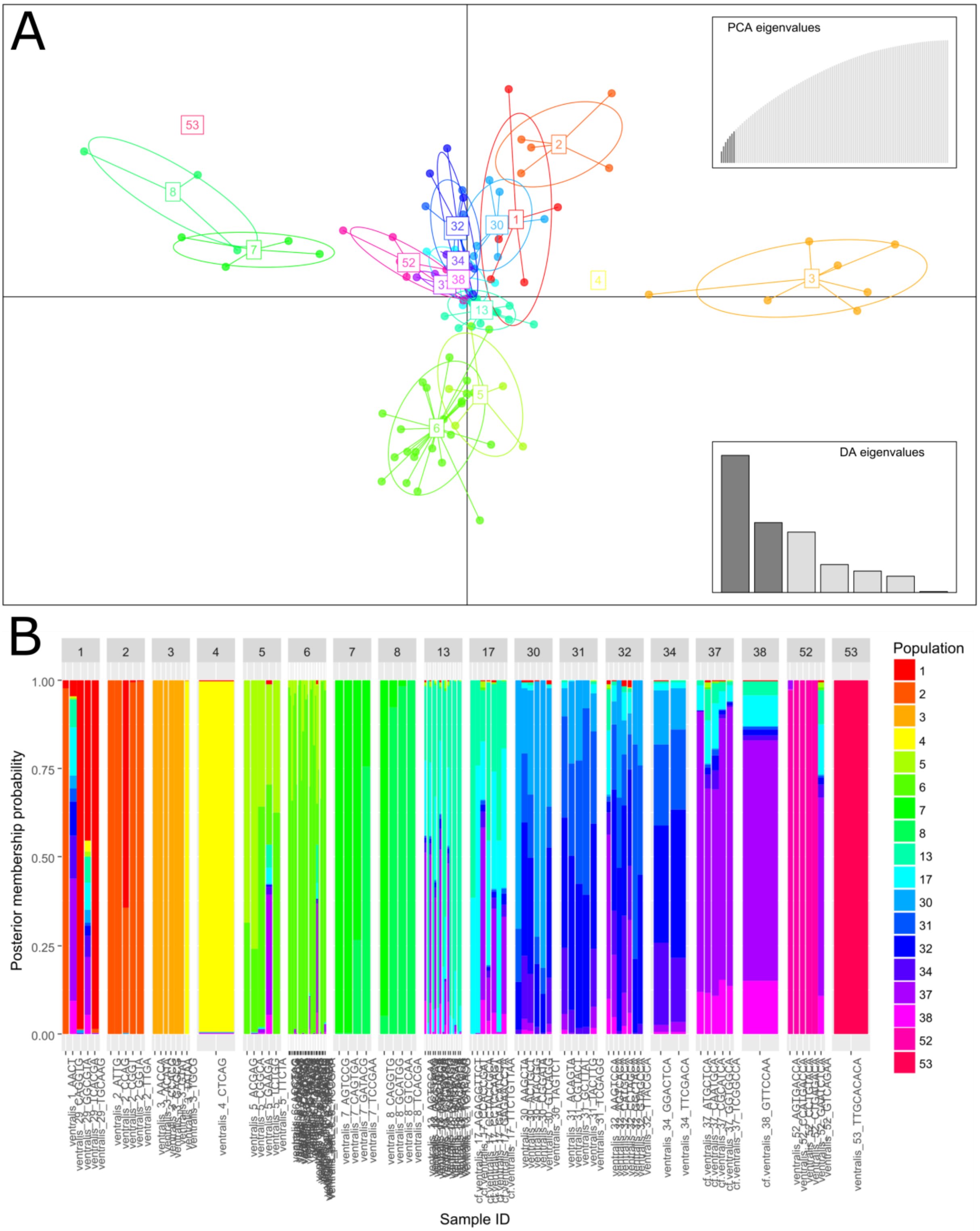

Figure 7. DAPC and resulting posterior membership probabilities for 110 specimen of *O. ventralis* sampled from 19 pooled localities, based on 24 703 SNPs. A) DAPC where the horizontal axis represents the first linear discriminant, the vertical axis

represents the second linear discriminant. Individuals (dots) are coloured and clustered based on the locality they were sampled. Localities are specified by their number. B) Compoplot displaying the posterior membership probabilities of each *O. ventralis* individual as a bar, cumulatively stacked from 0 to 1. Individual sample IDs for each bar and number identifiers for each locality are provided below and above the compoplot. Colour codes for the localities are shown on the right and correspond to the colours used in the DAPCs.

SNPs that contributed heavily to differentiation between lake-wide populations of *O. nasuta* resulted in a list of 20 SNPs along the first linear discriminant (LD 1) and 65 SNPs along the second linear discriminant (LD 2). All of these 85 SNPs were situated in an exon, distributed on 21 of the 23 larger scaffolds, and *SNPdat* was able to provide a gene name for 67 of these SNPs. GO term names classified as 'biological process' were successfully retrieved from *Ensembl* for 19 gene names. Annotations are summarised in Table 2.

LD	Scaffold	SNP Position	Gene name	*Ensembl* Gene Stable ID	Biological process
1	NC_031965.1	3816947	LOC100702482		
2	NC_031965.1	6960780	LOC100698156		
2	NC_031965.1	15093988	galns	ENSONIG00000003051	metabolic process
2	NC_031965.1	15094124	galns	ENSONIG00000003051	metabolic process
1	NC_031965.1	23266835	csmd1		
2	NC_031966.1	4823425	shisa3		
2	NC_031966.1	15392675	enox2		
2	NC_031966.1	18494726	alg13		
2	NC_031966.1	29402931	LOC100702720		
1	NC_031967.1	286304	LOC100694381		
2	NC_031969.1	5456200	hlf		
2	NC_031969.1	16736585	dnaaf5	ENSONIG00000005166	
2	NC_031969.1	28996534	tef		
2	NC_031969.1	33845990	LOC100701914		
2	NC_031970.1	25996491	raly		
2	NC_031970.1	30568791	iars	ENSONIG00000014366	translation tRNA aminoacylation for protein translation aminoacyl-tRNA metabolism for translational fidelity isoleucyl-tRNA aminoacylation regulation of sprouting angiogenesis
2	NC_031971.1	2524868	LOC100707400		
1	NC_031971.1	14396092	LOC100710254		
2	NC_031971.1	22941831	dnajb14	ENSONIG00000017488	
2	NC_031972.1	3753987	LOC100694717		
2	NC_031972.1	15856729	LOC100707779		
2	NC_031973.1	4342156	tbkbp1	ENSONIG00000018224	
2	NC_031973.1	14858282	dnmbp	ENSONIG00000016351	regulation of Rho protein signal transduction regulation of molecular function intracellular signal transduction kidney development cilium assembly
2	NC_031973.1	23990637	cacna1g	ENSONIG00000019655	regulation of ion transmembrane transport transmembrane transport ion transport calcium ion transport calcium ion transmembrane transport
1	NC_031974.1	9351842	LOC100694123		

2	NC_031975.1	30920042	fez1	ENSONIG00000015790	
2	NC_031977.1	11007576	LOC100698136		
2	NC_031977.1	19698623	LOC100692736		
2	NC_031977.1	19707115	rgmb	ENSONIG00000013691	BMP signalling pathway
1	NC_031977.1	28075470	nsd3	ENSONIG00000014418	methylation histone lysine methylation
2	NC_031977.1	37665411	anxa3	ENSONIG00000014003	negative regulation of catalytic activity
2	NC_031977.1	37678642	fras1	ENSONIG00000014013	cell communication endoderm development epithelial structure maintenance fin development fin morphogenesis
2	NC_031978.1	6638982	LOC100710194		
2	NC_031978.1	8534812	dlg5		
2	NC_031978.1	19025922	LOC100706044		
2	NC_031979.1	26493814	veph1	ENSONIG00000005544	
2	NC_031979.1	26493814	ptx3		
2	NC_031980.1	5890325	LOC100701144		
1	NC_031980.1	8899314	xrn2	ENSONIG00000018720	nucleic acid phosphodiester bond hydrolysis RNA phosphodiester bond hydrolysis, exonucleolytic nucleobase-containing compound metabolic process mRNA processing
2	NC_031980.1	8907027	xrn2	ENSONIG00000018720	nucleic acid phosphodiester bond hydrolysis RNA phosphodiester bond hydrolysis, exonucleolytic nucleobase-containing compound metabolic process mRNA processing
1	NC_031980.1	10106150	pex7	ENSONIG00000018773	
2	NC_031980.1	11389746	LOC100695926		
2	NC_031980.1	16860772	rev3l	ENSONIG00000011378	translesion synthesis
1	NC_031980.1	22885704	atp13a3	ENSONIG00000002627	cation transport
2	NC_031980.1	27275899	LOC100693372		
2	NC_031981.1	21211662	nup107	ENSONIG00000017999	nucleocytoplasmic transport mRNA export from nucleus nuclear pore complex assembly
2	NC_031981.1	25653066	lmnb2		
1	NC_031981.1	31095477	uchl5	ENSONIG00000010950	ubiquitin-dependent protein catabolic process protein deubiquitination proteolysis cranial skeletal system development
2	NC_031981.1	37551395	LOC100690600		
2	NC_031982.1	4445269	LOC100711149		
1	NC_031983.1	10158882	LOC100706370		
1	NC_031984.1	13127923	LOC100703401		
1	NC_031984.1	13127933	LOC100703401		
2	NC_031984.1	15846150	LOC100695192		
2	NC_031985.1	8578715	igf2bp3		
1	NC_031985.1	14951610	itgb8	ENSONIG00000010860	cell adhesion integrin-mediated signalling pathway cell-matrix adhesion
2	NC_031986.1	8490480	fryl	ENSONIG00000003053	heart development
2	NC_031986.1	8490486	fryl	ENSONIG00000003053	heart development
1	NC_031986.1	22768195	LOC100697713		
2	NC_031986.1	39209794	LOC100691786		
2	NC_031987.1	10803332	plekhm3	ENSONIG00000017099	intracellular signal transduction
1	NC_031987.1	19198868	LOC100697712		
1	NC_031987.1	20047376	gpr156		
2	NC_031987.1	22774114	hnrnpa3		
2	NW_017615942.1	28654	LOC100697640		
2	NW_017616076.1	43601	LOC100701157		

2	NW_017616076.1	43673	LOC100701157

Table 2. Annotation of genes containing the SNPs that contribute the most to the genomic discrimination between lake-wide populations of *O. nasuta*. The first column indicates along which linear discriminant a particular SNP influenced the outcome of the DAPC. The second and third column provide the exact position of SNPs in the genome. The fourth and fifth column provide names and identifiers for the gene in which a given SNP is situated. The last column provides information about the biological process a gene is involved in.

4.2.3 Genome-wide differentiation patterns within the four Ophthalmotilapia *species*

Weir and Cockerham F_{ST}-values (Figure 8) ranged from 0 to 0.65 for the pair of *O. boops* populations that were considered sympatric, and from 0 to 0.76 for the allopatric populations. F_{ST}-values for both pairs are spread fairly similar across the genome. For *O. nasuta*, F_{ST}-values ranged from 0 to 0.42 for the sympatric pair. The allopatric pair ranged from 0 to 1 in F_{ST}-values. For both manhattan plots, F_{ST}-values are spread rather homogeneous across the genome, but SNPs compared between allopatric *O. nasuta* populations clearly reach much higher degrees of fixation. Weir and Cockerham F_{ST}-values for the sympatric *O. heterodonta* localities ranged from 0 to 0.86. Along the genome, some SNPs on scaffolds NC_031983 and NC_031987 might appear to have a somewhat higher degree of differentiation than would be expected based on other parts of that genome. The allopatric population pair had F_{ST}-values ranging from 0 to 1. In strong contrast to the sympatric *O. heterodonta* manhattan plot, a large portion of SNPs showcases high degrees of differentiation. This differentiation is distributed evenly along the genome. For the sympatric pair of *O. ventralis* populations, F_{ST}-values ranged from 0 to 0.64. The allopatric pair had F_{ST}-values in the range of 0 to 0.92. Overall fixation is higher than in the sympatric counterpart, but not as pronounced as in the allopatric pairs of *O. nasuta* and *O. heterodonta.* F_{ST}-values are distributed fairly even across the genome for both the sympatric and the allopatric pairs of *O. ventralis.*

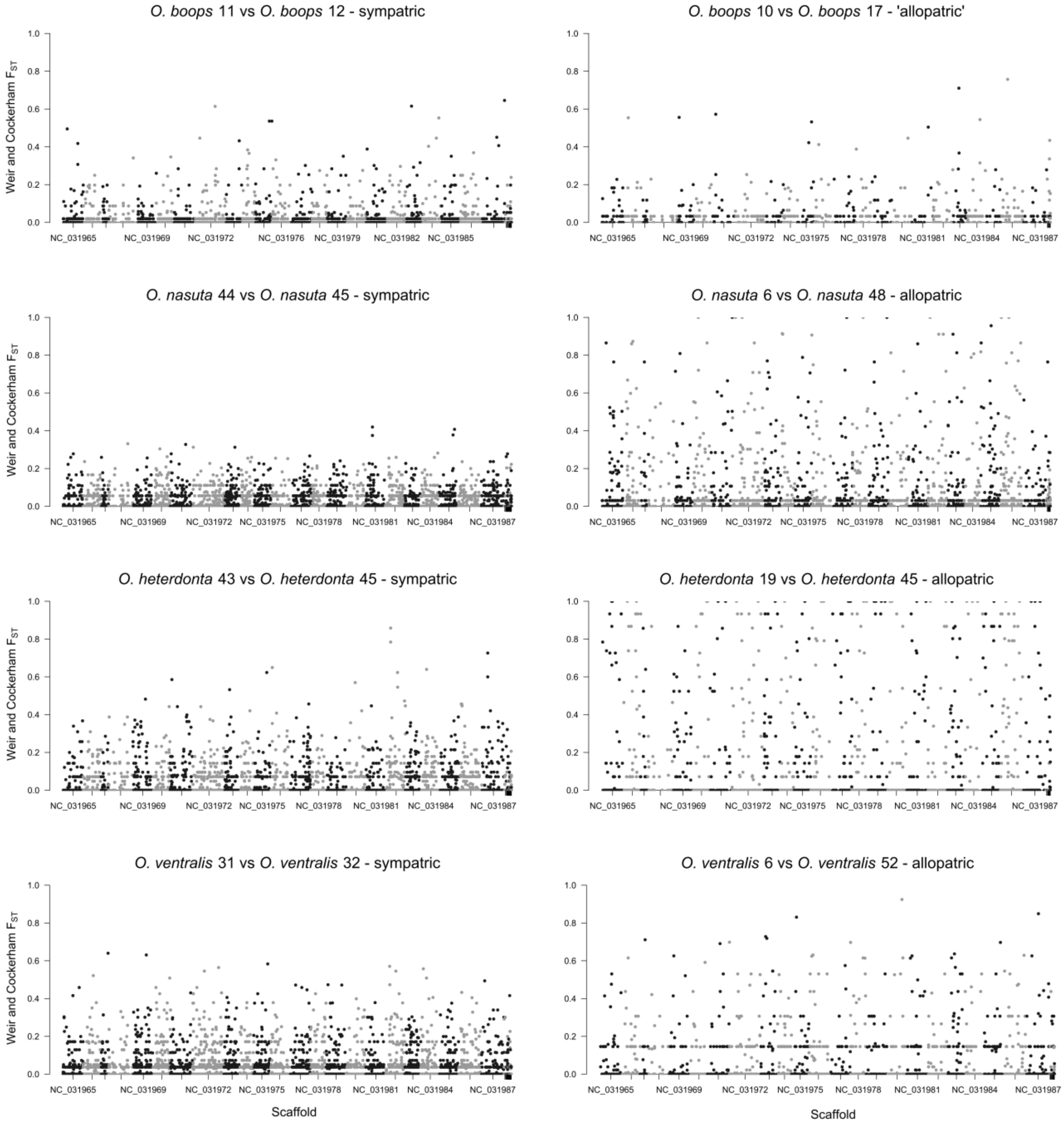

Figure 8. Manhattan plots for intraspecific pairs of populations of four *Ophthalmotilapia* species. Plots in the left side column are considered sympatric, plots in the right side column are considered allopatric. Weir and Cockerham F_{ST}-values are indicated on the vertical axis and range from 0 to 1. The scaffolds on which the SNPs are situated are given along the horizontal axis. Dots represent SNPs. SNPs alternate from grey to black along consecutive scaffolds for clarity.

4.2.4 Interspecific population genomic analysis of three sympatric Ophthalmotilapia *species*

For the DAPC of the sympatric region (Figure 9A), being the eastern side of the southern subbasin, the first six PCs were retained based on an a-score computation. This setting conserved 38.2% of variance.

Three linear discriminants were retained. LD 1 discriminated between each of the sympatric species, with *O. nasuta* occupying the central position. LD 2 also discriminated between all three species, with *O. boops* situated in between *O. nasuta* and *O. ventralis.* Individuals of *O. boops* and *O. ventralis* are mostly spread out along LD 1, whilst individuals of *O. nasuta* are spread out along LD 2. Individuals assigned to *O.* cf. *ventralis* cluster with *O. ventralis* along both axes. Strikingly, a number of individuals classified as *O. nasuta* associate completely with the *O. ventralis* cluster. A similar observation is made for a number of *O. boops* individuals, which associate strongly with the *O. nasuta* cluster.

Posterior membership probabilities computed across the species level (Figure 9B) reveal that most individuals show little admixture, indicating distinct and species-specific genomic compositions. However, a number of *O. boops* individuals show large degrees of admixture, especially with *O. nasuta.* Oddly enough, some *O. boops* individuals are attributed a *O. nasuta* membership probability of almost 100%. The same pattern is true for a number of *O. nasuta* individuals, which are assigned *O. ventralis* probabilities of close to 100%. Apart from this observation, varying degrees of two-directional admixture can be found between all species. Furthermore, individuals assigned to *O.* cf. *ventralis* cannot be distinguished reliably from *O. ventralis* individuals based on SNP composition and vice versa.

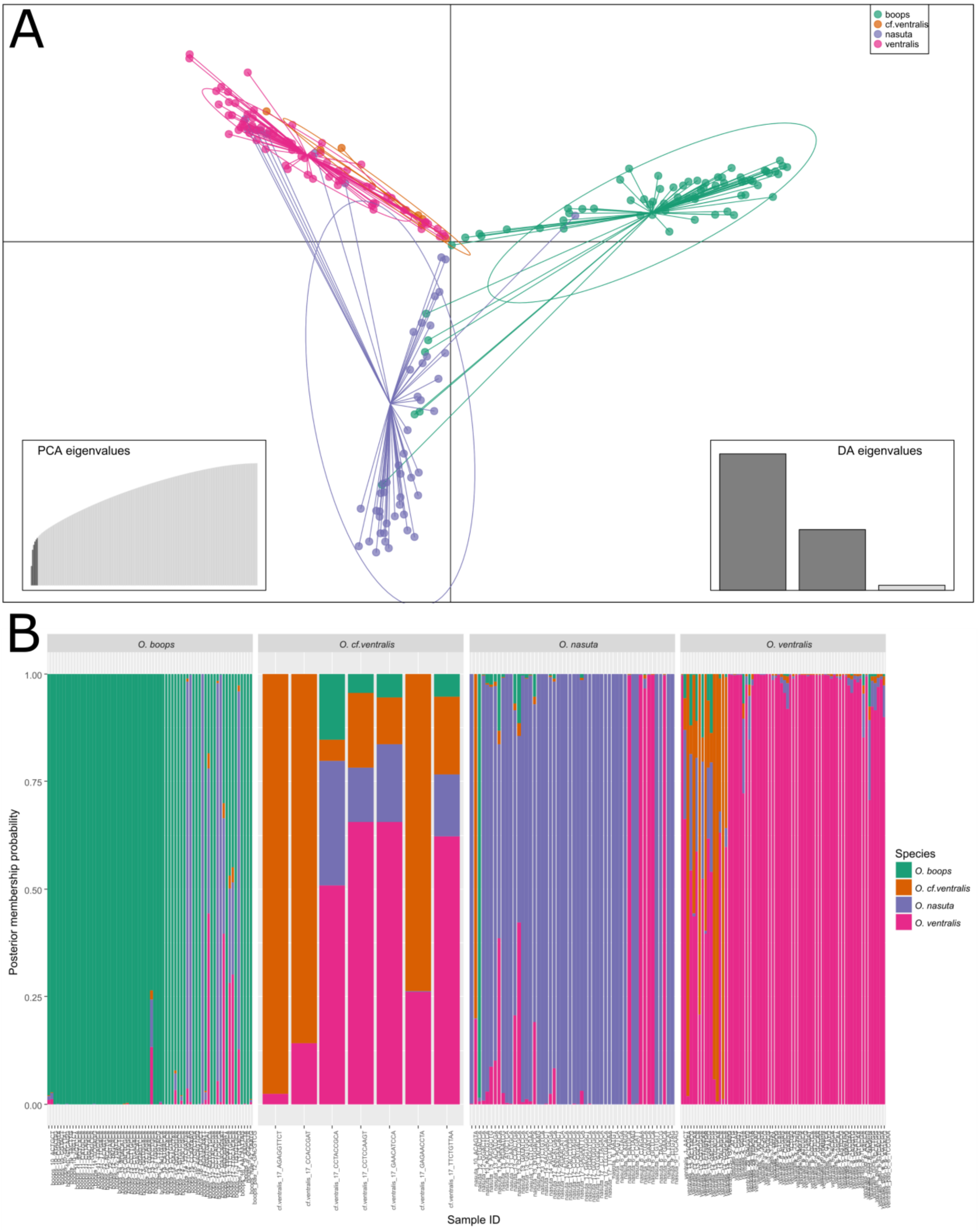

Figure 9. DAPC and resulting posterior membership probabilities for 197 specimen sampled throughout the eastern side of the southern subbasin (region ES), based on 31 914 SNPs. A) DAPC where the horizontal axis represents the first linear discriminant, the vertical axis represents the second linear discriminant. Individuals (dots) are coloured and clustered based on their taxonomic affinity. B) Compoplot displaying the posterior membership probabilities of each individual of the

sympatric region as a bar, cumulatively stacked from 0 to 1. Individual sample IDs for each bar are provided below the compoplot. Colour codes for the species are shown on the right and correspond to the colours used in the DAPC.

Selection of SNPs that contributed heavily to differentiation between sympatric populations of different species in the ES geographic region resulted in a list of 32 SNPs along the first linear discriminant (LD 1) and 25 SNPs along the second linear discriminant (LD 2). All 57 SNPs were flagged as exonic, and 48 of these were successfully assigned a gene name. GO term names classified as 'biological process' were retrieved from *Ensembl* for thirteen gene names. These annotations are summarised in Table 3. In the same manner, a list of SNPs was compiled for a subset similar to the sympatric subset, but with all the individuals excluded that do not cluster with their morphologically predetermined species assignment (Table S1). Retrieval of SNPs with high loading values resulted in a list of 12 SNPs along LD 1 and 31 SNPs along LD 2. Again, all of these 43 SNPs were flagged as exonic, and 36 SNPs were assigned a gene name. GO term names representing biological processes were attained for eight gene names (Table 4). Seventeen gene names are shared between Table 3 and Table 4.

LD	**Scaffold**	**SNP Position**	**Gene name**	***Ensembl* Gene Stable ID**	**Biological process**
1	NC_031965.1	7459775	luzp2		
2	NC_031965.1	20522236	myo1e		
1	NC_031965.1	24285761	galnt2	ENSONIG00000017082	protein glycosylation
1	NC_031966.1	9037290	LOC100695038		
2	NC_031966.1	13076962	tenm3		
1	NC_031966.1	19499415	mid2		
1	NC_031966.1	28504494	LOC102076836		
1	NC_031968.1	7299037	gal3st3	ENSONIG00000013030	glycolipid biosynthetic process
1	NC_031969.1	8584909	asic2	ENSONIG00000000822	ion transport sodium ion transport sodium ion transmembrane transport
1	NC_031970.1	13310020	srgap2	ENSONIG00000019295	signal transduction positive regulation of GTPase activity
2	NC_031970.1	23332662	LOC100696627		
2	NC_031970.1	23736512	nisch		
1	NC_031971.1	20333824	ctnna2	ENSONIG00000014573	cell adhesion brain morphogenesis
2	NC_031972.1	15615614	zfc3h1		
2	NC_031973.1	10977056	LOC100694933		
1	NC_031974.1	20047139	mllt10	ENSONIG00000014431	canonical Wnt signalling pathway intestinal epithelial structure maintenance
2	NC_031975.1	25382616	LOC100698500		
1	NC_031976.1	11534940	iglon5		
1	NC_031976.1	12072561	LOC100705345		
1	NC_031976.1	12072561	LOC109204190		
1	NC_031977.1	3005076	LOC100697308		
1	NC_031978.1	213182	pcsk2	ENSONIG00000000347	proteolysis

2	NC_031978.1	14464059	LOC100708959		
2	NC_031978.1	19055577	LOC100706044		
1	NC_031978.1	20030315	LOC100701755		
1	NC_031979.1	16432204	LOC100707847		
2	NC_031979.1	25006831	kirrel3		
1	NC_031979.1	35209679	tbx2	ENSONIG00000010454	regulation of transcription, DNA-templated transcription, DNA-templated
2	NC_031980.1	11419592	LOC100696183		
2	NC_031980.1	17111027	tab2	ENSONIG00000011396	heart development chordate embryonic development convergent extension involved in gastrulation
2	NC_031981.1	35819737	ror1	ENSONIG00000008814	protein phosphorylation transmembrane receptor protein tyrosine kinase - signalling pathway
1	NC_031981.1	37426264	camk1d		
2	NC_031983.1	6792913	LOC102081083		
2	NC_031983.1	9865227	LOC109195984		
1	NC_031984.1	13120967	LOC100703401		
2	NC_031985.1	16198714	wasf2	ENSONIG00000002236	actin cytoskeleton organisation
1	NC_031985.1	23589132	LOC109196657		
1	NC_031985.1	23589132	samd12		
2	NC_031986.1	934013	arhgef18		
2	NC_031986.1	23360579	LOC100701849		
2	NC_031987.1	17544813	cmss1	ENSONIG00000006678	
2	NC_031987.1	17544813	filip1l	ENSONIG00000012470	
1	NC_031987.1	17831508	il1rapl1	ENSONIG00000012110	signal transduction cytokine-mediated signalling pathway
1	NC_031987.1	38082501	LOC100691527		
1	NW_017613886.1	893277	cdc42bpa		
2	NW_017615946.1	155692	birc6	ENSONIG00000008016	apoptotic process protein ubiquitination regulation of cytokinesis
2	NW_017615949.1	240241	dmpk		
2	NW_017615949.1	240297	dmpk		

Table 3. Annotation of genes containing the SNPs that contribute the most to the genomic discrimination between sympatric species of *Ophthalmotilapia.* The first column indicates along which linear discriminant a particular SNP influenced the outcome of the DAPC. The second and third column provide the exact position of SNPs in the genome. The fourth and fifth column provide names and identifiers for the gene in which a given SNP is situated. The last column provides information about the biological process a gene is involved in.

LD	Scaffold	SNP Position	Gene name	Gene stable ID	Biological process
1	NC_031965.1	7459775	luzp2		
2	NC_031965.1	14461814	LOC100710496		
1	NC_031965.1	32133117	neto2		
1	NC_031969.1	8584909	asic2	ENSONIG00000000822	ion transport ligand-gated sodium channel activity sodium ion transport sodium ion transmembrane transport
2	NC_031969.1	29435217	LOC100690815		
2	NC_031971.1	20333824	ctnna2	ENSONIG00000014573	cell adhesion

					brain morphogenesis
2	NC_031971.1	23818820	tbc1d9		
2	NC_031972.1	14674714	immp2l		
1	NC_031972.1	15615614	zfc3h1		
2	NC_031973.1	23925685	cacna1g	ENSONIG00000019655	regulation of ion transmembrane transport transmembrane transport ion transport calcium ion transport calcium ion transmembrane transport calcium ion import
2	NC_031974.1	20047139	mllt10	ENSONIG00000014431	canonical Wnt signalling pathway intestinal epithelial structure maintenance
2	NC_031975.1	22467169	LOC100690123		
2	NC_031976.1	11534940	iglon5		
2	NC_031978.1	213182	pcsk2	ENSONIG00000000347	proteolysis
2	NC_031978.1	625017	LOC100694992		
2	NC_031978.1	14657805	macrod2		
1	NC_031978.1	19055577	LOC100706044		
1	NC_031978.1	20030315	LOC100701755		
1	NC_031979.1	16432204	LOC100707847		
2	NC_031980.1	23092407	ankrd13c	ENSONIG00000002649	
2	NC_031981.1	37426264	camk1d		
2	NC_031983.1	8546152	slc8a3	ENSONIG00000019591	transmembrane transport cell communication calcium ion transport sodium ion transmembrane transport
2	NC_031983.1	9865227	LOC109195984		
1	NC_031984.1	13120967	LOC100703401		
2	NC_031984.1	31059679	LOC100693447		
2	NC_031985.1	14164154	dync1i1	ENSONIG00000010764	microtubule-based movement
2	NC_031985.1	17386165	LOC100703333		
2	NC_031985.1	23589132	LOC109196657		
2	NC_031985.1	23589132	samd12	ENSONIG00000013544	
2	NC_031986.1	20670830	agap1	ENSONIG00000003589	positive regulation of GTPase activity
2	NC_031986.1	27381775	LOC109197001		
2	NC_031987.1	32663117	LOC106097784		
2	NC_031987.1	32663117	thsd7b		
1	NW_017613886.1	893277	cdc42bpa		
1	NW_017615949.1	240297	dmpk		
2	NW_017615953.1	37177	LOC100697214		

Table 4. Annotation of genes containing the SNPs that contribute the most to the genomic discrimination between the main species clusters observed in Figure 13. All ambiguously clustering individuals were removed and SNP loading values were computed again. The first column indicates along which linear discriminant a particular SNP influenced the outcome of the DAPC. The second and third column provide the exact position of SNPs in the genome. The fourth and fifth column provide names and identifiers for the gene in which a given SNP is situated. The last column provides information about the biological process a gene is involved in.

5. Discussion

5.1 *Phylogenetic and phylogeographic analysis using mtDNA*

Mitochondrial haplotypes within the haplotype network are separated by few mutational steps (Figure 2). This suggests that barriers to gene flow are not very effective, and that a number of haplotypes have originated quite recently. Avise et al. (1987) describe five major phylogeographic patterns based on mtDNA, and our results can be categorised under what the authors describe as a continuous genetic divergence pattern with nested geographic distribution. Under this pattern, some haplotypes are geographically widespread (here generally corresponding with those with above-average connectedness) and allied haplotypes are more localised (here generally corresponding with those with one or two connections). This pattern implies that presumed ancestral haplotypes occur over a broad area, and that newly arisen mutations have not yet spread throughout the range of a species (Avise et al. 1987). In agreement with the study by Nevado et al. (2011), a significant number of haplotypes is shared between species. The authors of the previous study propose three possibilities for this observation regarding *Ophthalmotilapia*, and settle for hybridisation as the most likely cause, thereby dismissing both faulty identification of specimen and incomplete lineage sorting. Mitochondrial haplotypes shared between species is often reported (Rohwer et al. 2001, Grant and Grant 2008). A cluster of closely associated haplotypes specific to *O. nasuta* includes individuals from all regions, indicating occurrences of lake-wide gene flow within *O. nasuta*. However, two branches of three *O. nasuta* haplotypes each are restricted to a single respective region and separated from the bulk by multiple mutational steps. These lineages might be subjected to more persistent barriers to gene flow. In further agreement with the results obtained by Nevado et al. (2011), *O. nasuta* shares genetic variation with *O. boops*, *O. ventralis* and *O. heterodonta* in regions where *O. nasuta* individuals live in sympatry with each of the other *Ophthalmotilapia*. The authors concluded that this introgression was unidirectional into *O. nasuta*. We found that *O. boops* not only shares haplotypes with sympatric *O. nasuta*, but with sympatric *O. ventralis* and *O.* cf. *ventralis* as well, implying that hybridisation events between *Ophthalmotilapia spp*. not necessitate the involvement of *O. nasuta* and might tell a more complicated story. Remarkably, two haplotypes are shared between *O. boops* and *O. heterodonta*, which do not occur in sympatry and as such are more likely the result of either incomplete lineage sorting or an indirect result of hybridisation with a third species as vector (McDonald et al. 2008). Congruent with earlier remarks (Hanssens et al. 1999, Nevado et al. 2011, Konings 2014), *O. ventralis* and *O. heterodonta* individuals cannot be assigned to clearly demarcated species clusters. Concerning *O.* cf. *ventralis*, the small amount of individuals available cluster in two haplotypes that are separated by at least five mutational steps, offering no mitochondrial genetic support for their potential distinctness (Konings 2014).

5.2 *Population genomic analysis using genotyping-by-sequencing data*

5.2.1 *Intraspecific population genomic analysis of the four* Ophthalmotilapia *species*

Analysis of the SNP data reveals a distinctive population structure for each of the four species of *Ophthalmotilapia. O. boops* (Figure 4) is geographically confined to one continuous section of shoreline on the eastern banks of southern Lake Tanganyika. Analysing its population genomic structure hints at a slight break in continuity, dividing the section in a southern and northern subpopulation, possibly reflecting a habitat barrier. Genetic variation within these clusters is substantial, indicating considerable localised gene flow. Individuals from locality 14, right in between the two subpopulations, have genetically diverged from both. Not coincidentally, these individuals were sampled on an island (Nvuna Island) some distance off the coast. In the most southern subpopulation, localities show gradual admixture in function of distance, indicating gene flow mostly between neighbouring localities. This trend is less visible in the northern subpopulation, possibly due to lower sample numbers in these localities. Some individuals from Nvuna Island and from the northern subpopulation carry a large proportion of genomic variation typical of the southern population, while the reverse is not true, which could suggest that dispersal occurs mainly northwards. Since *O. boops* is confined to a single region, the scale of the mtDNA analysis is too coarse to make any comparisons with results from genomic DNA.

Genetic differentiation in *O. nasuta* (Figure 5) is clearly more pronounced between clusters than within, reflecting a high degree of structure. Localities along the eastern shore of the southern subbasin show complete admixture and therefore can be interpreted as one panmictic subpopulation, with extensive gene flow between all of its localities. Localities situated on the same side of the lake but spread out to the north of the panmictic subpopulation seem to be genetically isolated, as they differentiate on the DAPC and show almost no admixture. Two more subpopulations can be identified, both on the western shoreline of the northern subbasin. The more southern of the two consists of five localities that show gradual admixture. Although no admixture with the panmictic subpopulation can be detected, save one individual from locality 43, both subpopulations cluster together on the DAPC. A likely explanation could be that they are differentiated along linear discriminants other than the first and the second. These findings are compatible with those based on mtDNA, where a cluster of associated haplotypes was identified along with two isolated branches, here represented by the admixed populations and the isolated populations respectively. Strangely, localities 19, 35 and 43 contain an individual that was attributed practically 0% membership probability for its respective locality. Immigration of these particular individuals from another locality stands as the most evident explanation. However, the foreign specimen in locality 19 originates from locality 35 and vice versa, and the distance between locality 19 and 35 is huge, from the central eastern region to the west side of the southern subbasin. Since no similar signs of gene flow are present in other individuals from those localities, it seems more likely that the two relevant samples got swapped somehow. Regardless of these individuals, our results arguably agree with the finding that shallower shorelines in Lake Tanganyika can harbour more genetic diversity in rock-dwelling cichlids than steeper shorelines (Nevado et al. 2013). The authors propose that fluctuations in the water level of Lake Tanganyika have resulted in periodic strong episodes of migration between populations that are otherwise isolated, allowing genetic diversity to spread (Nevado et al. 2013), and it is suggested that this could result in higher total diversity once they become isolated again (Wakeley and Aliacar 2001). Some of the most isolated *O. nasuta* populations are sampled in shallow localities (23, 25-27, 35, 40, 47-51) and conversely, the most admixed populations are found in the steepest localities (9-17, 43-46). This indicates that structuring was influenced by historic lake level changes. However, not all localities followed this pattern (22, 6, 43) and sample sizes of some localities is minimal. Moreover, exact assessment of what constitutes a deep rather than a shallow locality is needed.

Deduced from small amounts of variation between genotypes of the same locality compared to differentiation between localities, *O. heterodonta* (Figure 6) is subjected to a very high degree of genomic population structuring. Localities generally do not cluster together, with the exception of those situated in the southern west end of the northern subbasin. Even in those localities, only a small portion of individuals show some admixture, and we can conclude from the posterior membership probabilities that localities of *O. heterodonta* are strongly segregated, and gene flow is uncommon between them. This conclusion was to be expected, as variations in mitochondrial haplotypes of *O. heterodonta* are abundant, and haplotypes are not shared between many individuals, indicating that individuals cluster in small, isolated populations. Whether these results contradict the pattern observed by Nevado et al. (2013) and outlined above for *O. nasuta* is not clear. Part of the small amounts of admixture that was detected does occur in deep localities (43-45), but also in shallow localities (19-20).

Variation within clusters of *O. ventralis* (including *O.* cf. *ventralis*; Figure 7) is rather high and most localities cluster together, but some degree of structuring can be observed. Individuals along the lower east side of Lake Tanganyika seem to be structured in smaller genetic associations, with localities clustering on their own or in pairs. Little discrimination is made in the DAPC between individuals from both lakesides of the northern end of the lower subbasin, possibly echoing historic connectedness from a time of low lake stand when these lakesides were united as one northern shoreline. Surprisingly, localities from the southern west side are part of this same cluster. Admixture is least prominent in the lower east region, supporting its higher degree of structuring. The lower west region (localities 30-34), in contrast, is mostly admixed, indicating extensive gene flow between these localities. This region and the localities in the northern end of the distribution range all show signs of genomic introgression, in all directions. Some lower eastern individuals carry a lot of genetic material from the western side (in localities 1, 5, 6) suggesting dispersal from west to east. Comparison of these results with the constructed mitochondrial haplotype network is complex, but in general, results from both analyses support the claim that populations along the east side of Lake Tanganyika are more structured than those along the west side. Since populations from both sides largely reside in shallow localities, but are nonetheless subjected to different degrees of population structuring, the SNP data does not support the negative correlation between steepness genetic variation as observed in *O. nasuta* and Nevado et al. (2013). This confirms the findings of Sefc et al. (2007), who found little differentiation between populations of *O. ventralis* in the most southern part of the west side of Lake Tanganyika.

Derycke et al. (2018) compiled a list of 36 genes that have been linked to fish behaviour and physiological networks in previous studies, and found that a number of these genes were upregulated in the brain of female *O. ventralis* and *O. nasuta* individuals that were kept in a social setting with conspecific females. None of these 36 genes is found to be among the genes associated with high-profile *O. nasuta* SNPs (Table 2). In the same study, the authors identified the top five upregulated GO biological processes for each of six macroanatomical brain regions of the two species after mating experiments. One of these biological processes, cell adhesion, which they found upregulated in the optic tectum of both species, is also found in this study for *O. nasuta* (Table 2). Comparison of the processes retrieved for *O. nasuta* in this study to the broad functional groupings of processes that Malinsky et al. (2015) deemed potentially important in the adaptive divergence of two crater lake cichlid populations, suggests the following processes are relevant to the divergence of *O. nasuta* populations: fin development, fin morphogenesis, regulation of sprouting angiogenesis, cranial skeletal system development, and regulation of Rho protein signal transduction. The first four are related to the functional grouping 'morphogenesis', the fifth process is related to 'sensory systems' (Malinsky et al. 2015). The processes under the umbrella 'morphogenesis' are consistent with morphological differentiation. The regulation of Rho protein signal transduction is involved in the functioning of rhodopsin, a light-sensitive receptor protein found in the retina (Malinsky et al. 2015). The potential role of visual perception in cichlid speciation is well known (Seehausen et al. 2008).

5.2.2 Genome-wide differentiation patterns within the four Ophthalmotilapia *species*

A common statistical measure of genetic differentiation between populations is F_{ST}, which compares the variation in allele frequencies between populations to the variance within populations (Weir and Cockerham 1984). Comparisons of allopatric *Ophthalmotilapia* populations revealed a pattern of consistently high F_{ST}-values across the genome (Figure 8), while comparisons of sympatric *Ophthalmotilapia* display mainly low F_{ST}-values. Some high F_{ST}-values in the sympatric *O. heterodonta* manhattan plot can be observed, but they can hardly be interpreted as an island of differentiation and overall, F_{ST}-values of SNPs within *Ophthalmotilapia* populations are homogeneously distributed across the genome. In the context of the debate regarding 'islands/continents of differentiation', these results comply most convincingly with the mechanism of continents, i.e. a genome-wide reduction in the effective migration rate of alleles caused by divergent selection on traits with a polygenic architecture. The observed lack of locally differentiated genomic regions could be statistically supported with an outlier analysis, which involves the identification of genomic regions with F_{ST}-values above the maximum levels that can be generated under neutral coalescent simulations. Additional predictions can be tested to determine the validity of putative islands, such as a measurement of absolute sequence divergence (Malinsky et al. 2015). It should be noted that alternative explanations can be offered for an observed lack of differentiation islands. Apart from methodological failure to detect islands, they could be absent because epistatic interactions are at play (Wolf and Ellegren 2017) or simply due to a lack of divergent selection. Regardless of the island discussion, SNPs compared between allopatric populations clearly reach far higher degrees of fixation (with the exception of *O. boops*, which lacks truly allopatric populations). This indicates that the allopatric localities are more diverged than the sympatric pairs. Notably, none of the SNPs that approximate fixation in the allopatric O. nasuta populations match the SNPs that were identified as strong discriminators between lake-wide O. nasuta populations (Table 2). The SNPs differentiating between allopatric O. ventralis populations do not reach fixation, in contrast to those of allopatric O. nasuta and O. heterodonta populations, suggesting more gene flow between allopatric O. ventralis populations. Since allopatric populations of O. nasuta and O. heterodonta are disconnected over larger distances than those of O. ventralis and especially of O. boops, this indicates that geographic distance poses a barrier to gene flow within species.

5.2.3 Interspecific population genomic analysis of three sympatric Ophthalmotilapia *species*

In large part, the sympatric *Ophthalmotilapia* samples cluster within their respective species (Figure 9A), and although considerable variation exists within these species clusters, species are clearly segregated from one another. Numerous samples from all three sympatric species pools carry a proportion of DNA that is attributed to another species (Figure 9B). These findings are consistent with the results obtained from mitochondrial DNA (Figure 2), where *O. boops, O. nasuta* and *O. ventralis* are mostly separated as well, and genomic admixture is in concordance with the amount of haplotype sharing deduced from the mitochondrial data. As Nevado et al. (2011) concluded, admixture patterns of *O. nasuta* individuals indicate introgression into *O. nasuta* from both *O. boops* and *O. ventralis*. The authors based their conclusions mainly on mitochondrial DNA, and regarded the results of their gene flow analysis with nine nuclear microsatellites, which retrieved signals of gene flow between all four *Ophthalmotilapia,* as inconclusive. However, given the admixture we observe based on SNP data between each of the three species that occur in the lower eastern subbasin, their nuclear DNA results are partly replicated in this study. This would imply that gene flow within the genus *Ophthalmotilapia* does not follow a unidirectional pattern into *O. nasuta* but occurs in both directions between at least three species. Introgression of *O. nasuta* and *O. ventralis* material into *O. boops*, and of *O. ventralis* into *O. nasuta*, appears to be the most extensive. Based on SNP genotyping of seven individuals of *O.* cf. *ventralis*, no clear genetic distinction can be made that would set *O.* cf. *ventralis* apart from *O. ventralis*. Inexplicably, seven individuals of *O. nasuta* and six individuals of *O. boops* are assigned to, respectively, *O. nasuta* and *O. ventralis* with membership probabilities nearing 100%. Since we assume that the SNPs are distributed all over the genome, this would imply that almost all of the genome is typical of another species. It seems very improbable that an individual would have the phenotypic characteristics of one species and the genotype of another, so the most parsimonious explanation would be either misidentification of the specimen, or swapping of samples during the sequencing process. Studying the voucher specimen, if available, coupled with resequencing, would be required to solve this matter.

Incomplete lineage sorting can be suggested as an alternative explanation to hybridisation for shared genetic variation, and the distinction is not easily made (Barton 2001, Gante et al. 2016, Alter et al. 2017, Zhou et al. 2017). Nevado et al. (2011) provides ample arguments in favour of hybridisation, but these are rooted in results from mitochondrial data and do not automatically hold for results obtained through a genome-wide analysis. In the scenario of incomplete lineage sorting, variation from a common ancestor shared by multiple lineages is expected to be distributed randomly with respect to geographic patterns (Barton 2001, Alter et al. 2017). Since populations of *Ophthalmotilapia* spp. are geographically structured, hybridisation between different species should occur mostly in regions where those species live in sympatry. Therefore, the proportion of shared variation is expected to be lower in allopatric populations of different species, which can be tested by calculating posterior membership probabilities for allopatric populations of the three species here compared in sympatry. Another method by which genome-wide hybridisation signals in *Ophthalmotilapia* spp. could be verified is by employing ABBA-BABA tests, also known as Patterson’s D statistic, to test for excess of shared allelic variants under strict bifurcation (Green et al. 2010, Gante et al. 2016). Although aforementioned tests should be performed to conclude with certainty that the observed shared SNP variation between sympatric *Ophthalmotilapia* spp. can be attributed to hybridisation, the following arguments provide reasonable grounds to consider hybridisation more likely than incomplete lineage sorting: 1) hybridisation is the preferred explanation based on mtDNA and microsatellites (Nevado et al. 2011); 2) hybridisation has actually been observed in lab aquaria (Kéver et al. 2018); 3) proportions of heterospecific DNA can be high (> 25%), indicating continuous introgression (Hertwig et al. 2009, Nevado et al. 2011); and 4) most *O. heterodonta* populations and some *O. ventralis* and *O. nasuta* populations share no variation with conspecific populations, indicating that lineage sorting has been completed among those populations. Since population divergence events are more recent than speciation events, it follows that lineage sorting should be completed among species as well.

The SNPs that differentiate between sympatric *Ophthalmotilapia* spp. (Table 3 & 4) were evaluated in light of the same studies as those that differentiate between lake-wide populations of *O. nasuta* (Table 2). Again, none of the retrieved genes corresponds to those that are listed by Derycke et al. (2018) based on literature. As in the *O. nasuta* subset, cell adhesion is listed among the sympatric results, a process that was also differentially expressed in the brain of O. nasuta and O. ventralis females. In addition, intestinal epithelial structure maintenance is detected by Derycke et al. (2018) as well. Considering the functional categories established by Malinsky et al. (2015), the following processes could be involved in upholding the *Ophthalmotilapia* species barriers: brain morphogenesis, chordate embryonic development and convergent extension involved in gastrulation, all of which might be related to morphogenesis. A recent study also discussed the immune system, and inflammatory response and cytokine activity in particular, as a functional category involved in cichlid radiation (Malinsky et al. 2017). The biological processes actin cytoskeleton organisation, cytokine-mediated signalling pathway and regulation of cytokinesis (Table 3) are in accordance with this statement.

6. Conclusions

Population differentiation was high in *O. nasuta* and *O. heterodonta*, but less so in *O. ventralis* and *O. boops*. Patterns of population structure derived from genomic and mitochondrial data were compatible, and the capacity for genome-wide data to provide more detail was illustrated. Apart from *O. ventralis,* these patterns appear to correspond with a trend for increased genetic diversity in populations inhabiting shallow areas, which suggests that historic lake level changes in Lake Tanganyika helped shape present-day population structures. We found no indication of genetic islands of differentiation along the genome of allopatric *Ophthalmotilapia* populations. Analysis of genome-wide data from congeneric individuals living in sympatry, representing three different species of

Ophthalmotilapia, revealed shared genetic variation between each of those three species. This result can most likely be attributed to hybridisation. If so, interspecific gene flow seems to occur in both directions between each of the three sympatric *Ophthalmotilapia* species, rather than unidirectional into *O. nasuta*. Annotation of highly divergent SNPs recovered some biological processes that are known to be involved in cichlid speciation, such as morphogenesis, vision, and cytokine activity.

7. Supplementary Material

Supplementary material is distributed digitally.

Table S1: IDs of individuals analyzed, and assignment into subsets.

Table S2: PCR protocols.

Figure S1: a-score distributions for DAPC.

8. Acknowledgements

This manuscript is the final product of my master's dissertation, and I thank the University of Ghent and the Royal Belgian Institute of Natural Sciences for providing me with the opportunity, for funding and for support in a variety of other ways. I also thank the members and partners of the GENBAS project for presenting me a conceptual and practical framework in which to conduct this study, and for supplying me with data. I further thank Sofie Derycke, Maarten Van Steenberge and Erik Verheyen for their supervision and scientific advice. Maarten is also thanked for his technical support along with helpful comments and teaching, and special thanks go to Sofie for her tutoring, recommendations and comments on earlier versions of this manuscript, as well as her continual support in general.

9. References

Agarwala, R., T. Barrett, J. Beck, D. A. Benson, C. Bollin, E. Bolton, D. Bourexis, J. R. Brister, S. H. Bryant, K. Canese, M. Cavanaugh, C. Charowhas, K. Clark, I. Dondoshansky, M. Feolo, L. Fitzpatrick, K. Funk, L. Y. Geer, V. Gorelenkov, A. Graeff, W. Hlavina, B. Holmes, M. Johnson, B. Kattman, V. Khotomlianski, A. Kimchi, M. Kimelman, M. Kimura, P. Kitts, W. Klimke, A. Kotliarov, S. Krasnov, A. Kuznetsov, M. J. Landrum, D. Landsman, S. Lathrop, J. M. Lee, C. Leubsdorf, Z. Y. Lu, T. L. Madden, A. Marchler-Bauer, A. Malheiro, P. Meric, I. Karsch-Mizrachi, A. Mnev, T. Murphy, R. Orris, J. Ostell, C. O'Sullivan, V. Palanigobu, A. R. Panchenko, L. Phan, B. Pierov, K. D. Pruitt, K. Rodarmer, E. W. Sayers, V. Schneider, C. L. Schoch, G. D. Schuler, S. T. Sherry, K. Siyan, A. Soboleva, V. Soussov, G. Starchenko, T. A. Tatusova, F. Thibaud-Nissen, K. Todorov, B. W. Trawick, D. Vakatov, M. Ward, E. Yaschenko, A. Zasypkin, K. Zbicz, and N. R. Coordinators. 2018.

Database resources of the National Center for Biotechnology Information. Nucleic Acids Research **46**:D8-D13.

Alter, S. E., J. Munshi-South, and M. L. J. Stiassny. 2017. Genomewide SNP data reveal cryptic phylogeographic structure and microallopatric divergence in a rapids-adapted clade of cichlids from the Congo River. Molecular Ecology **26**:1401-1419.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific Phylogeography - the Mitochondrial-DNA Bridge between Population-Genetics and Systematics. Annual Review of Ecology and Systematics **18**:489-522.

Bandelt, H.-J., P. Forster, and A. Röhl. 1999. Median-joining networks for inferring intraspecific phylogenies. Molecular Biology and Evolution **16**:37-48.

Barton, N. H. 2001. The role of hybridization in evolution. Molecular Ecology **10**:551-568.

Barton, N. H., and G. M. Hewitt. 1989. Adaptation, Speciation and Hybrid Zones. Nature **341**:497-503.

Brower, A. V. Z. 2011. Hybrid speciation in Heliconius butterflies? A review and critique of the evidence. Genetica **139**:589-609.

Chorowicz, J. 2005. The East African rift system. Journal of African Earth Sciences **43**:379-410.

Cohen, A. S., K. E. Lezzar, J. J. Tiercelin, and M. Soreghan. 1997. New palaeogeographic and lake-level reconstructions of Lake Tanganyika: Implications for tectonic, climatic and biological evolution in a rift lake. Basin Research **9**:107-132.

Cohen, A. S., M. J. Soreghan, and C. A. Scholz. 1993. Estimating the Age of Formation of Lakes - an Example from Lake Tanganyika, East-African Rift System. Geology **21**:511-514.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and G. P. A. Grp. 2011. The variant call format and VCFtools. Bioinformatics **27**:2156-2158.

Deperi, S. I., M. E. Tagliotti, M. C. Bedogni, N. C. Manrique-Carpintero, J. Coombs, R. F. Zhang, D. Douches, and M. A. Huarte. 2018. Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. Plos One **13**.

Derycke, S., L. Kever, K. Herten, K. Van den Berge, M. Van Steenberge, J. Van Houdt, L. Clement, P. Poncin, E. Parmentier, and E. Verheyen. 2018. Neurogenomic Profiling Reveals Distinct Gene Expression Profiles Between Brain Parts That Are Consistent in Ophthalmotilapia Cichlids. Frontiers in Neuroscience **12**.

Doran, A. G., and C. J. Creevey. 2013. Snpdat: Easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. Bmc Bioinformatics **14**.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. Plos One **6**.

Feder, J. L., S. P. Egan, and P. Nosil. 2012. The genomics of speciation-with-gene-flow. Trends in Genetics **28**:342-350.

Fitzpatrick, B. M., J. A. Fordyce, and S. Gavrilets. 2009. Pattern, process and geographic modes of speciation. Journal of Evolutionary Biology **22**:2342-2347.

Fryer, G., T. D. Iles, and C. M. Yonge. 1972. The cichlid fishes of the great lakes of Africa : their biology and evolution. Edinburgh : Oliver and Boyd.

Gante, H. F., M. Matschiner, M. Malmstrom, K. S. Jakobsen, S. Jentoft, and W. Salzburger. 2016. Genomics of speciation and introgression in Princess cichlid fishes from Lake Tanganyika. Molecular Ecology **25**:6143-6161.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. arXiv preprint.

Genner, M. J., P. Nichols, G. Carvalho, R. L. Robinson, P. W. Shaw, A. Smith, and G. F. Turner. 2007. Evolution of a cichlid fish in a Lake Malawi satellite lake. Proceedings of the Royal Society B-Biological Sciences **274**:2249-2257.

Grant, B. R., and P. R. Grant. 2008. Fission and fusion of Darwin's finches populations. Philosophical Transactions of the Royal Society B-Biological Sciences **363**:2821-2829.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. W. Zhai, M. H. Y. Fritz, N. F. Hansen, E. Y. Durand, A. S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prufer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Hober, B. Hoffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, and S. Paabo. 2010. A Draft Sequence of the Neandertal Genome. Science **328**:710-722.

Haffer, J. 2008. Hypotheses to explain the origin of species in Amazonia. Brazilian Journal of Biology **68**:917-947.

Hanssens, M., J. Snoeks, and E. Verheyen. 1999. A morphometric revision of the genus Ophthalmotilapia (Teleostei, Cichlidae) from Lake Tanganyika (East Africa). Zoological Journal of the Linnean Society **125**:487-512.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, and G. O. Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research **32**:D258-D261.

Herten, K., M. S. Hestand, J. R. Vermeesch, and J. K. J. Van Houdt. 2015. GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. Bmc Bioinformatics **16**.

Hertwig, S. T., M. Schweizer, S. Stepanow, A. Jungnickel, U. R. Bohle, and M. S. Fischer. 2009. Regionally high rates of hybridization and introgression in German wildcat populations (Felis silvestris, Carnivora, Felidae). Journal of Zoological Systematics and Evolutionary Research **47**:283-297.

Hewitt, G. M. 2004. Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **359**:183-195.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. 2002. The Ensembl genome database project. Nucleic Acids Research **30**:38-41.

Immler, S., and M. Taborsky. 2009. Sequential polyandry affords post-mating sexual selection in the mouths of cichlid females. Behavioural Ecology and Sociobiology **63**:1219-1230.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics **24**:1403-1405.

Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. Bmc Genetics **11**.

Jombart, T. C., C. 2015. A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 2.1.0. Page 43. Imperial College London.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody, S. White, E. Birney, S. Searle, J. Schmutz, J. Grimwood, M. C. Dickson, R. M. Myers, C. T. Miller, B. R. Summers, A. K. Knecht, S. D. Brady, H. Zhang, A. A. Pollen, T. Howes, C. Amemiya, P. Broad Institute Genome Sequencing, T. Whole Genome Assembly, E. S. Lander, F. Di Palma, K. Lindblad-Toh, and D. M. Kingsley. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484**:55.

Kéver, L., E. Parmentier, S. Derycke, E. Verheyen, J. Snoeks, M. Van Steenberge, and P. Poncin. 2018. Limited possibilities for prezygotic barriers in the reproductive behaviour of sympatric Ophthalmotilapia species(Teleostei, Cichlidae).

Knaus, B. J., and N. J. Grunwald. 2017. VCFR: a package to manipulate and visualise variant call format data in R. Molecular Ecology Resources **17**:44-53.

Koblmuller, S., W. Salzburger, B. Obermuller, E. Eigner, C. Sturmbauer, and K. M. Sefc. 2011. Separated by sand, fused by dropping water: habitat barriers and fluctuating water levels steer the evolution of rock-dwelling cichlid populations in Lake Tanganyika. Molecular Ecology **20**:2272-2290.

Konings, A. 2014. Featherfins in their natural habitat. Cichlid Press.

Kumar, S., G. Stecher, and K. Tamura. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Molecular Biology and Evolution **33**:1870-1874.

Lande, R. 1981. MODELS OF SPECIATION BY SEXUAL SELECTION ON POLYGENIC TRAITS. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences **78**:3721-3725.

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods **9**:357-U354.

Larson, W. A., L. W. Seeb, M. V. Everett, R. K. Waples, W. D. Templin, and J. E. Seeb. 2014. Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (Oncorhynchus tshawytscha). Evolutionary Applications **7**:355-369.

Leigh, J. W., and D. Bryant. 2015. POPART: full-feature software for haplotype network construction. Methods in Ecology and Evolution **6**:1110-1116.

Lezzar, K. E., J. J. Tiercelin, M. DeBatist, A. S. Cohen, T. Bandora, P. VanRensbergen, C. LeTurdu, W. Mifundu, and J. Klerkx. 1996. New seismic stratigraphy and Late Tertiary history of the North Tanganyika Basin, East African Rift system, deduced from multichannel and high-resolution reflection seismic data and piston core evidence. Basin Research **8**:1-28.

Magoc, T., and S. L. Salzberg. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics **27**:2957-2963.

Malinsky, M., R. J. Challis, A. M. Tyers, S. Schiffels, Y. Terai, B. P. Ngatunga, E. A. Miska, R. Durbin, M. J. Genner, and G. F. Turner. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science **350**:1493-1498.

Malinsky, M., H. Svardal, A. M. Tyers, E. A. Miska, M. J. Genner, G. F. Turner, and R. Durbin. 2017. Whole genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Page 33.

McDonald, D. B., T. L. Parchman, M. R. Bower, W. A. Hubert, and F. J. Rahel. 2008. An introduced and a native vertebrate hybridize to form a genetic bridge to a second native species. Proceedings of the National Academy of Sciences of the United States of America **105**:10837-10842.

McGlue, M. M., K. E. Lezzar, A. S. Cohen, J. M. Russell, J. J. Tiercelin, A. A. Felton, E. Mbede, and H. H. Nkotagu. 2008. Seismic records of late Pleistocene aridity in Lake Tanganyika, tropical East Africa. Journal of Paleolimnology **40**:635-653.

Michel, A. P., S. Sim, T. H. Q. Powell, M. S. Taylor, P. Nosil, and J. L. Feder. 2010. Widespread genomic divergence during sympatric speciation. Proceedings of the National Academy of Sciences of the United States of America **107**:9724-9729.

Nevado, B., V. Fazalova, T. Backeljau, M. Hanssens, and E. Verheyen. 2011. Repeated Unidirectional Introgression of Nuclear and Mitochondrial DNA Between Four Congeneric Tanganyikan Cichlids. Molecular Biology and Evolution **28**:2253-2267.

Nevado, B., S. Koblmuller, C. Sturmbauer, J. Snoeks, J. Usano-Alemany, and E. Verheyen. 2009. Complete mitochondrial DNA replacement in a Lake Tanganyika cichlid fish. Molecular Ecology **18**:4240-4255.

Nevado, B., S. Mautner, C. Sturmbauer, and E. Verheyen. 2013. Water-level fluctuations and metapopulation dynamics as drivers of genetic diversity in populations of three Tanganyikan cichlid fish species. Molecular Ecology **22**:3933-3948.

Papadopoulou, A., and L. L. Knowles. 2015. Genomic tests of the species-pump hypothesis: Recent island connectivity cycles drive population divergence but not speciation in Caribbean crickets across the Virgin Islands. Evolution **69**:1501-1517.

Peart, C. R., K. K. Dasmahapatra, and J. J. Day. 2018. Contrasting geographic structure in evolutionarily divergent Lake Tanganyika catfishes. Ecology and Evolution **8**:2688-2697.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rohwer, S., E. Bermingham, and C. Wood. 2001. Plumage and mitochondrial DNA haplotype variation across a moving hybrid zone. Evolution **55**:405-422.

Rossiter, A. 1995. The Cichlid Fish Assemblages of Lake Tanganyika: Ecology, Behaviour and Evolution of its Species Flocks. Pages 187-252 *in* M. Begon and A. H. Fitter, editors. Advances in Ecological Research. Academic Press.

Rundle, H. D., and P. Nosil. 2005. Ecological speciation. Ecology Letters **8**:336-352.

Seehausen, O., Y. Terai, I. S. Magalhaes, K. L. Carleton, H. D. J. Mrosso, R. Miyagi, I. van der Sluijs, M. V. Schneider, M. E. Maan, H. Tachida, H. Imai, and N. Okada. 2008. Speciation through sensory drive in cichlid fish. Nature **455**:620-U623.

Seehausen, O., J. J. M. vanAlphen, and F. Witte. 1997. Cichlid fish diversity threatened by eutrophication that curbs sexual selection. Science **277**:1808-1811.

Sefc, K. M., S. Baric, W. Salzburger, and C. Sturmbauer. 2007. Species-specific population structure in rock-specialized sympatric cichlid species in Lake Tanganyika, East Africa. Journal of Molecular Evolution **64**:33-49.

Sturmbauer, C. 1998. Explosive speciation in cichlid fishes of the African Great Lakes: a dynamic model of adaptive radiation. Journal of Fish Biology **53**:18-36.

Sturmbauer, C., S. Baric, W. Salzburger, L. Ruber, and E. Verheyen. 2001. Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes. Molecular Biology and Evolution **18**:144-154.

Turner, S. D. 2014. qqman: an R package for visualising GWAS results using Q-Q and manhattan plots. bioRxiv.

Verheyen, E., L. Ruber, J. Snoeks, and A. Meyer. 1996. Mitochondrial phylogeography of rock-dwelling cichlid fishes reveals evolutionary influence of historical lake level fluctuations of Lake Tanganyika, Africa. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences **351**:797-805.

Via, S., and J. West. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Molecular Ecology **17**:4334-4345.

Vines, T. H., S. C. Kohler, A. Thiel, I. Ghira, T. R. Sands, C. J. MacCallum, N. H. Barton, and B. Nurnberger. 2003. The maintenance of reproductive isolation in a mosaic hybrid zone between the fire-bellied toads Bombina bombina and B-variegata. Evolution **57**:1876-1888.

Wakeley, J., and N. Aliacar. 2001. Gene genealogies in a metapopulation. Genetics **159**:893-905.

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population-Structure. Evolution **38**:1358-1370.

Winkelmann, K., L. Ruber, and M. J. Genner. 2017. Lake level fluctuations and divergence of cichlid fish ecomorphs in Lake Tanganyika. Hydrobiologia **791**:21-34.

Wolf, J. B. W., and H. Ellegren. 2017. Making sense of genomic islands of differentiation in light of speciation. Nature Reviews Genetics **18**:87-100.

Zhou, Y., L. Duvaux, G. Ren, L. Zhang, O. Savolainen, and J. Liu. 2017. Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. Heredity **118**:211-220.

II. Dutch Summary

Om nieuwe inzichten te verwerven in de genetische grondslag van het speciatieproces, wordt steeds vaker genoomwijd onderzoek toegepast op populaties uit een natuurlijke omgeving. In deze studie gebruiken we SNP's, gegenotypeerd met gebruik van de *genotype-by-sequencing* methode, om populatiestructuur, patronen van *gene flow* en signalen van hybridisatie te onderzoeken in het geslacht *Ophthalmotilapia*, een klein geslacht van cichliden uit het Tanganyikameer. Er werden stalen ingezameld van de vier soorten doorheen hun respectievelijke verspreidingsgebieden. Dit laat toe om gedetailleerde conclusies te trekken omtrent de *gene flow* tussen zowel sympatrische als allopatrische populaties van dezelfde soort. Aangezien deze verspreidingsgebieden erg verschillen tussen soorten, maar deels overlappen, kan ook de *gene flow* tussen sympatrische soorten aan het licht gebracht worden. Bovendien maken SNP data het ook mogelijk om te verkennen hoe genetische variatie verdeeld is langsheen het genoom, en welke biologische processen verantwoordlijk zijn voor deze variatie. We tonen aan dat er meer populatiegenomische structuur aanwezig is in *O. nasuta* en *O. heterodonta* dan in *O. ventralis* en *O. boops*., en leggen voorzichtig de link met historische fluctuaties in het waterpeil van het Tanganyikameer. De resultaten die we verkrijgen voor intraspecifieke analyses, gebaseerd op genomische data, zijn in overeenstemming met eerdere conclusies die naar voor geschoven werden op basis van mitochondriaal onderzoek, met het verschil dat genomische data een gedetailleerder beeld leveren. Analyses tussen verschillende soorten geven echter een ander resultaat dan wat verwacht werd aan de hand van mitochondriale bevindingen: we vinden gedeelde genetische variatie tussen elk van drie soorten *Ophthalmotilapia* in een regio waar ze symptrisch voorkomen, wat de eerdere stelling dat *gene flow* zich in één richting voltrekt, met *O. nasuta* als ontvangende soort, aan het wankelen brengt. Onze resultaten wijzen eerder aan dat *gene flow* tussen elk van de sympatrische soorten plaatsvindt, en dat telkens in twee richtingen. Verder wenden we de SNP data aan om F_{ST}-waarden te berekenen voor sympatrische en allopatrische populaties van alle vier de soorten *Ophthalmotilapia,* en vinden geen bewijs voor de hypothese dat genetische variatie tussen divergerende populaties georganiseerd is in lokale pieken van differentiatie langsheen het genoom. Ten slotte identificeren we de SNPs die het meest voor genetische spreiding tussen populaties zorgen, en annoteren deze met het meest recente *Oreochromis niloticus* genoomassemblage. Een klein aantal van de geannoteerde SNP's ligt in genen die mogelijks betrokken zijn in biologische processen waarvan reeds geweten is dat ze een rol kunnen spelen in de speciatie van cichliden.

III. Appendix

A1	**Folder containing SeqMan Assembly Projects** This folder contains D-loop sequences from the 32 *O. boops* samples that I prepared for sequencing (extraction, PCR, PCR clean-up, gel electrophoresis) in the lab at the RBINS. Not all extractions or PCRs were successful, yielding 23 sequences that I checked by eye in SeqMan.
A2	**FASTA of partial D-loop sequences** This FASTA file contains the 490 partial D-loop sequences that were used for the mtDNA analysis, along with six outgroup sequences from GenBank. Sequences are aligned and trimmed.
A3	**FASTA of unique partial D-loop sequences** I used FaBox (http://users-birc.au.dk/biopv/php/fabox/) to collapse all 490 sequences to 137 unique haplotypes, to see whether it would give me better bootstrap values on my phylogenetic tree (see A4).
A4	**Maximum Likelihood tree in NEXUS format** Bootstrap values were very low when I constructed a ML tree and did not improve by using only unique haplotypes, which is why this tree is not included in the manuscript. The tree was constructed in MEGA and is saved in NEXUS format. When opened in FigTree, taxa are colour-coded. Since the 'import colour scheme' function in FigTree is not actually implemented, I coloured the taxa using a script from GitHub (https://github.com/acorg/figtree-recolor).
A5	**Settings and summary statistics for the ML tree** As outputted from MEGA.
A6	**Species-based traits file for PopART** To colour a haplotype network in PopART based on some trait (e.g. taxonomy), a traits file needs to be provided along with the sequence data. PopART would not accept my traits files, and I sidestepped the problem by adding a TRAITS block to the sequence data (A8-9).
A7	**Region-based traits file for PopART** See A6.
A8	**PopART input file with species as trait** To construct a haplotype network coloured by species, mitochondrial sequences were inputted as a NEXUS file, with a TRAITS block added that contains the species information.
A9	**PopART input file with geographic region as trait** Same sequences as A8 but with a TRAITS block that contains geographic information.
A10	**VCFtools pipeline** Documentation of filter steps used to filter the SNP data. I worked with the file 'freebayes.m15.q15.useduplicates.ploidy2_.vcf', which contained the first two GBS runs of the GENBAS project (384 individuals, 5,9 GB of data).
A11	**Missingness values** Excel file with the missingness values of the initial SNP dataset as outputted by VCFtools. Other tabs list the individuals that should be removed from the initial dataset to keep certain subsets.
A12	**SNPhylo pipeline** I tried constructing a phylogenetic tree from SNP data using the SNPhylo tool, but I was not able to solve a recurrent error and had to forfeit this analysis.
A13	**R script for DAPC with pooled localities** The R script that was used to make DAPCs and compoplots for each species, with pooled localities as cluster IDs.
A14	**R script for DAPC with geographic regions** The R script that was used to make DAPCs and compoplots for each species, with geographic regions as cluster IDs. These are the same regions that were used for the mtDNA haplotype network. Since using pooled localities proved more informative, the DAPCs with regional cluster IDs are not included in the manuscript.

A15	**R script for DAPC in the sympatric region** The R script that was used to make the DAPC with the three species that occur sympatrically in the ES region.
A16	**Geographic information for *O. boops*** This text file contains the information needed to group *O. boops* individuals based on sampling locality, and is required to run A13 and A14.
A17	**Geographic information for *O. nasuta*** This text file contains the information needed to group *O. nasuta* individuals based on sampling locality, and is required to run A13 and A14.
A18	**Geographic information for *O. heterodonta*** This text file contains the information needed to group *O. heterodonta* individuals based on sampling locality, and is required to run A13 and A14.
A19	**Geographic information for *O. ventralis*** This text file contains the information needed to group *O. ventralis* individuals based on sampling locality, and is required to run A13 and A14.
A20	**Taxonomic information for the sympatric region (ES)** This text file contains the information needed to group individuals from the sympatric area by species, and is required to run A15.
A21	**Taxonomic information for the sympatric region (ES) (2)** This text file contains the information needed to group individuals from the sympatric area by species, and is required to run A15. The difference with A20 is that this file does not contain the individuals that cluster almost 100% with another species.
A22	**Manhattan script** The R script to make manhattan plots. This script requires the text files in the A24 folder that contain F_{ST}-values.
A23	**Shell script to compute pairwise F_{ST}-values** This unix shell script runs a repetitive VCFtools command to calculate F_{ST}-values between sympatric and allopatric populations of each species. It requires text files from A24, and provides output that is required for A22. Output files are not readily usable by the manhattan R script, but need some regex expression transformations to fit the format expected by the manhattan script. Negative values are set to zero.
A24	**Folder with manhattan-related text files** This folder contains text files with the individuals in each population, required for A23, and polished output files of A23 with F_{ST}-values, required for A22.
A25	**SNP annotation script for *O. nasuta*** The R script that generates the input files needed by SNPdat to annotate SNPs that discriminate between lake-wide populations of *O. nasuta*.
A26	**SNP annotation script for the sympatric region** The R script that generates the input files needed by SNPdat to annotate SNPs that discriminate between sympatric species of *Ophthalmotilapia*.
A27	**SNP annotation script for the sympatric region (2)** The R script that generates the input files needed by SNPdat to annotate SNPs that discriminate between sympatric species of *Ophthalmotilapia*. The difference with A26 is that this analysis does not contain the individuals that cluster almost 100% with another species.
A28	**SnpEff pipeline** Brief documentation of a failed attempt to annotate SNPs with the SnpEff tool (http://snpeff.sourceforge.net/). In the end it turned out that I was using the wrong genome assembly for annotation, but by then I was already using SNPdat.
A29	**SNPdat pipeline** Detailed pipeline on the use of SNPdat.
A30	**Shell script to execute SNPdat annotation** This unix shell script uses a VCF file and text files generated by A25-27 as input files in VCFtools, to pass VCF files to SNPdat that only contain the SNPs of interest. It then outputs annotated SNP data.

A31	**BioMart pipeline** Procedure to retrieve information on biological processes for the genes discovered by SNPdat.
A32	**Part of the annotated SNPs for *O. nasuta*** This excel file serves as an example in A31.
A33	**QGIS files** Shapefiles and coordinates to produce the sampling map, included in the manuscript, with QGIS. (https://www.qgis.org/nl/site/).