MOLECULAR ECOLOGY

Molecular Ecology (2012) doi: 10.1111/mec.12083

Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes

I. KELLER,*† C. E. WAGNER,*† L. GREUTER,*† S. MWAIKO,*† O. M. SELZ,*† A. SIVASUNDAR,*†‡ S. WITTWER*† and O. SEEHAUSEN*†

*Department of Fish Ecology and Evolution, Center of Ecology, Evolution and Biochemistry, EAWAG Swiss Federal Institute of Aquatic Science and Technology, Seestrasse 79, CH-6047, Kastanienbaum, Switzerland, †Department of Aquatic Ecology and Macroevolution, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, CH-3012, Bern, Switzerland

Abstract

Adaptive radiations are an important source of biodiversity and are often characterized by many speciation events in very short succession. It has been proposed that the high speciation rates in these radiations may be fuelled by novel genetic combinations produced in episodes of hybridization among the young species. The role of such hybridization events in the evolutionary history of a group can be investigated by comparing the genealogical relationships inferred from different subsets of loci, but such studies have thus far often been hampered by shallow genetic divergences, especially in young adaptive radiations, and the lack of genome-scale molecular data. Here, we use a genome-wide sampling of SNPs identified within restriction site–associated DNA (RAD) tags to investigate the genomic consistency of patterns of shared ancestry and adaptive divergence among five sympatric cichlid species of two genera, *Pundamilia* and *Mbipia*, which form part of the massive adaptive radiation of cichlids in the East African Lake Victoria. Species pairs differ along several axes: male nuptial colouration, feeding ecology, depth distribution, as well as the morphological traits that distinguish the two genera and more subtle morphological differences. Using outlier scan approaches, we identify signals of divergent selection between all species pairs with a number of loci showing parallel patterns in replicated contrasts either between genera or between male colour types. We then create SNP subsets that we expect to be characterized to different extents by selection history and neutral processes and describe phylogenetic and population genetic patterns across these subsets. These analyses reveal very different evolutionary histories for different regions of the genome. To explain these results, we propose at least two intergeneric hybridization events (between *Mbipia* spp. and *Pundamilia* spp.) in the evolutionary history of these five species that would have lead to the evolution of novel trait combinations and new species.

Keywords: divergent selection, East African cichlids, *Mbipia*, outlier scan, *Pundamilia*, RAD sequencing, speciation

Received 1 July 2012; revision received 29 August 2012; accepted 7 September 2012

Introduction

A major pattern in the evolution of species diversity is the heterogeneity of speciation rates in space and time. Bursts of speciation are often associated with adaptive radiation, whereby many new phenotypically diverse

Correspondence: I. Keller, Fax: +41 (0)58 765 21 68; E-mail: irene.keller@eawag.ch

‡Present address: National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India

species arise from a common ancestor in short succession in response to ecological opportunity (Simpson 1953; Schluter 2000; Gavrilets & Losos 2009). One of the most outstanding recent radiations is the cichlid fishes of Lake Victoria, where more than 500 genetically distinct and phenotypically diverse species have arisen within <100 000 years, and possibly as recently as in the last 15 000 years (Johnson *et al.* 1996; Seehausen 2006; Bezault *et al.* 2011).

The demonstration of genetic distinctiveness among these many young species has been a major challenge and, until very recently, it had proven impossible to reconstruct phylogenetic relationships within the Lake Victoria radiation (Samonte *et al.* 2007; Bezault *et al.* 2011; Konijnendijk *et al.* 2011). Incomplete lineage sorting, phenotypic diversification without speciation and historical episodes of hybridization have all been invoked to explain these difficulties. Recent work found that phenotypic diversification in this system is generally associated with speciation (Bezault *et al.* 2011), but hybridization between species has clearly occurred (Seehausen *et al.* 2008). It has been proposed that occasional, spatially confined hybridization between differentiated species in an adaptive radiation may provide genetic fuel for further bouts of adaptive diversification (Grant & Grant 1992; Seehausen 2004), and strong support for this hypothesis has recently come from the radiation of *Heliconius* butterflies (*Heliconius* Genome Consortium 2012). The lack of phylogenetic resolution among the cichlid species of Lake Victoria has made studying the evolutionary consequences of hybridization in this group difficult until now. Very recent work, using short read sequence data from a very large number of loci, has for the first time demonstrated both sharp genetic boundaries between Lake Victoria cichlid species, and well-supported relationships between them (Wagner *et al.* 2012). However, the estimated phylogenetic relationships in this recent work did not match morphology-based taxonomy, and several morphologically based genera were rendered para- and polyphyletic. Here, we use SNPs generated from next-generation sequencing data to ask whether mismatches between morphology and phylogeny are because of parallel evolution of morphologies or because of a mosaic evolutionary history.

Gene flow because of hybridization is often thought to hinder adaptive divergence (e.g. Garant *et al.* 2007), and the origin and persistence of differentiated populations in parapatry or sympatry (Felsenstein 1981). However, sometimes gene flow may facilitate divergence by introducing novel genetic variation, which natural or sexual selection can then act upon. In several East African cichlid radiations, for example, there is repeated parallel divergence in male nuptial colouration (Allender *et al.* 2003). It is possible that similar male phenotypes evolved repeatedly and independently through novel mutation and that the observed parallelism is because of genetic or ecological constraints. Alternative and perhaps more likely scenarios, especially in very young radiations, suggest that similar phenotypes have the same genetic basis and that the same alleles were recruited repeatedly by natural or sexual selection from standing genetic variation, or that particular alleles were exchanged between more distantly related species through historical or recent hybridization.

Patterns of divergence between incipient and young species and the consistency of these patterns across the genome can provide valuable insights into the relative roles of different evolutionary forces. Genes involved in adaptive divergence and/or genetic incompatibilities between species (or loci linked to such genes), for example, are expected to show unusually high levels of allele sorting among species compared with the genomic average (Wu 2001; Beaumont 2005). If gene flow is globally restricted, we expect that markers under divergent selection could transiently show elevated genetic differentiation, but that selective and neutral loci would support the same or mutually consistent genealogical relationships. If gene flow is restricted only within narrow islands of differentiation, on the other hand, the population structure inferred from loci underlying adaptation may reflect the similarity or dissimilarity of selection pressures while, at neutral loci, it would reflect the patterns of ancestry and/or current gene flow (e.g. Wilding *et al.* 2001; Egan *et al.* 2008). Mismatched phylogenetic signals may be particularly pronounced if introgression within islands of differentiation is, or was at some point in the past, facilitated by selection in some populations or species but not in others.

Investigations into the heterogeneity of genomic divergence are becoming easier in nonmodel organisms because of the advent of next-generation sequencing technologies. In particular, these approaches allow a much higher genomic resolution as the number of markers that can be screened increases by orders of magnitude compared with traditional molecular methods. Here, we sequence a reduced representation library of RAD tags (Baird *et al.* 2008). By focusing on the regions flanking specific restriction sites, the complexity of the DNA fragment pool is reduced which, in turn, increases the read depth at each locus. By combining many individually barcoded samples into a single sequencer lane, it is possible to simultaneously detect and genotype thousands of SNPs from across the genome. The RAD-seq approach has been successfully used to investigate the heterogeneity of genomic divergence between different stickleback ecotypes (marine vs.

freshwater: Hohenlohe *et al.* 2010; stream vs. lake: Roesti *et al.* 2012). The method has also been used to identify many thousands of SNPs in species without an available reference genome (e.g. Amores *et al.* 2011; Barchi *et al.* 2011; Hohenlohe *et al.* 2011).

Our study species are sympatric populations of five haplochromine cichlids from the large adaptive radiation of Lake Victoria, East Africa. Speciation in this system has repeatedly been associated with parallel divergence of very similar male colour patterns in different genera, that is, in combination with different morphological traits (Seehausen & Van Alphen 1999; Seehausen *et al.* 1999, 2008). Here, we study two of these genera, each with a species with yellow–red male colouration and one or two with blue male colouration. Four of the species are formally described and grouped into two genera based on several morphological traits: *Pundamilia pundamilia*, *P. nyererei*, *Mbipia mbipi* and *M. lutea* (Seehausen *et al.* 1998). The fifth species is most similar to the two *Pundamilia* species based on morphological and dentition traits and has been provisionally named *Pundamilia* sp. 'pink anal fin' (Seehausen 1996).

We use SNPs identified from a *de novo* assembly of RAD tag sequences to investigate patterns of adaptive divergence and the genomic consistency of phylogenetic relationships among these species using a combination of outlier scan approaches, analyses of population structure and phylogenetic methods. We are specifically interested in assessing the proportion of loci with elevated levels of differentiation between species and in comparing population genetic structure and phylogenetic relationships at such loci with those estimated from the rest of the genome. For doing so, we create SNP subsets characterized by different levels of divergence across the five species and investigate whether the inferred population structure and phylogenetic relationships vary between these data subsets. Specifically, we ask how relationships between species with different generic assignment and different male colouration change between SNP subsets.

Material and methods

Study species

All samples for this study were collected at Makobe Island, an offshore island in southern Lake Victoria, Tanzania, characterized by high water transparency and a very species-rich cichlid community (Seehausen & Bouton 1997). We sampled ten males each from five species of the genera *Pundamilia* and *Mbipia* in 2010: *Pundamilia nyererei*, *P. pundamilia*, *P.* sp. 'pink anal fin', *Mbipia mbipi* and *M. lutea.* These species are among the 11 numerically most abundant species at this location, with *P. nyererei* and *M. mbipi* among the three most abundant species (Seehausen & Bouton 1998; O. Seehausen, unpublished data; Table 1). The two genera are distinguished based on a number of phenotypic traits including head and body morphology, dentition, scale morphology and squamation patterns (Seehausen *et al.* 1998). Both genera show parallel divergence into pairs of sister species with blue vs. yellow/red male nuptial colouration (*P. pundamilia* vs. *P. nyererei*, *M. mbipi* vs. *M. lutea*; Seehausen *et al.* 1998). *P.* sp. 'pink anal fin' is a third species with blue male colouration, but its phylogenetic position was unresolved at the outset of this

Table 1 Study species from Makobe Island, Lake Victoria

	Mbipia lutea	*Mbipia mbipi*	*Pundamilia* sp. 'pink anal fin'	*Pundamilia nyererei*	*Pundamilia pundamilia*
					
Abundance*	68 (11)	3811 (2)	872 (5)	4686 (1)	260 (6)
Male nuptial colouration	Yellow	Blue	Blue	Yellow and red dorsum	Blue
Depth distribution†	0–100	50–350	375–800	375–800	0–200
Tooth shape‡	Bicuspid	Bicuspid	Unicuspid	Unicuspid	Unicuspid
Feeding ecology	Benthic, algae scraper	Benthic, algae scraper	Bentholimnetic, zooplanktivorous	Bentholimnetic, zooplanktivorous	Benthic, insect larvae

*Number of individuals (and abundance rank) among 15 000 fish caught between 1995 and 2003 (O. Seehausen, unpublished).
†Depth range (in cm) containing >80% of the population.
‡Shape of the tooth crown in the outermost tooth row.

study. *Mbipia lutea* is entirely yellow on the flanks, while *P. nyererei* is yellow on the lower half but crimson red on the upper half of the flanks, the dorsum and the dorsal fin. Previous work has shown that the expression of red dorsal nuptial colouration in Lake Victoria cichlids requires the expression of yellow flank colouration (Seehausen *et al.* 1999; Magalhaes & Seehausen 2010). For this reason, we treat these two colour types as one category ('yellow') and compare it with the category of blue colouration.

Ecological and microhabitat differences between the five species are indicated in Table 1. The *P. pundamilia–P. nyererei* species pair has been particularly well studied, and it has been demonstrated that these two taxa are fully isolated biological species at Makobe Island (Seehausen *et al.* 2008; Seehausen 2009). The remaining three taxa are also phenotypically highly distinct, and all individuals can be unequivocally assigned to a given species. Behavioural work has demonstrated assortative mating between *M. mbipi* and *M. lutea* (Verzijden *et al.* 2008). Recent phylogenomic work has shown that all five species form well-supported, reciprocally monophyletic groups, while the two genera are not recovered as monophyletic groups within the 16 species from eight morphologically-based genera included in those analyses (Wagner *et al.* 2012). All fish were identified to species level based on phenotype by O. Seehausen and O. Selz and are vouchered in collections at EAWAG.

Molecular methods

DNA was extracted from finclips using a DNeasy Blood & Tissue kit (Qiagen) following the manufacturer's instructions. RAD libraries were prepared following the detailed protocol outlined in Etter *et al.* (2011) with some modifications as described below, using an *SbfI* high fidelity restriction enzyme (New England Biolabs). The two genera were sequenced in separate libraries together with other species, and each library contained between 52 and 60 individually barcoded fish. The barcodes were 6 bp long and differed by at least two bases. We used 750 pmol P1 adaptor (Microsynth AG) per 1 μg of digested genomic DNA and, after multiplexing, the sample was sheared in a Sonorex Super sonicator (Bandelin electronic GmbH & Co. KG) using five 30 s on and off cycles. The final amplification was carried out in two separate 50 μL PCRs per library each with 18 amplification cycles. The two aliquots were combined before the final size selection for fragments of *c.* 300–500 bp.

All libraries were sequenced on an Illumina HiSeq 2000 platform at Fasteris (Geneva, Switzerland). Two sequencer lanes were used for the initial sequencing of libraries containing the 50 individuals. Some individuals were replicated once (one *M. lutea*) or twice (all *Pundamilia* individuals) in additional sequencing lanes to increase the total number of reads.

Quality filtering and SNP calling

Reads without the complete *SbfI* recognition sequence were discarded from further analyses with our own python script. Using the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), all sequences were end-trimmed to a length of 90 bp, and reads containing one or more bases with a Phred quality score below 10 or more than 5% of the positions below 30 were discarded. The libraries were demultiplexed using the process_radtags program from the Stacks pipeline (Catchen *et al.* 2011). Single errors within the barcode were automatically corrected by the software. The final quality filtered and demultiplexed data set contained *c.* 116 million reads, each 84 base pairs in length. The number of reads per individual ranged between 436 K and 6690 K and was systematically higher in the three *Pundamilia* species because of the higher sequencing effort (see also Fig. S1, Supporting information of Wagner *et al.* 2012).

All reads from the 50 individuals were pooled and used for a *de novo* assembly in ustacks (Catchen *et al.* 2011). A 'stack' is a set of identical sequences in the terminology of this pipeline; several of these stacks may then be merged to form putative loci. We set a minimum stack depth of 50 reads (m) and excluded all stacks with coverage lower than this threshold. The maximum number of pairwise differences allowed between any two stacks within a locus was set to 2 (M parameter of ustacks). Because this parameter constrains the number of pairwise differences between stacks, the number of polymorphic sites at a locus can be higher than M when more than two stacks are merged. We disabled the deleveraging algorithm of ustacks and excluded putative loci with unusually high coverage (i.e. 'lumberjack stacks' of ustacks). Note that ustacks will treat stacks differing by an insertion–deletion polymorphism as separate loci.

The *de novo* assembly produced *c.* 136K loci. We used the consensus sequences from these loci as a 'reference' against which we mapped the quality filtered and demultiplexed reads from each individual separately using bowtie v. 0.12.7 (Langmead *et al.* 2009) allowing a maximum of 2 mismatches. All reads with more than one valid alignment were excluded.

Genotypes were called for all individuals together with the Unified Genotyper from the Genome Analysis Tool Kit (GATK) v.1.4-19, using the SNP genotype likelihood model (McKenna *et al.* 2010; DePristo *et al.*

2011). We considered only bases with a Phred quality score of at least 20 and used biallelic SNP calling as recommended for the current version of the program. Consequently, a maximum of two different bases is possible at each site.

From the vcf file containing all polymorphic sites, we created subsets containing all possible species pairs to produce a total of 11 files per assembly: one full data set with five species and 10 subsets with two species each. The following filtering steps were performed on each file: all SNP positions with a Phred quality score below 20 were excluded. Next, we recoded all individual genotypes with a quality score (i.e. GQ field of the vcf file) below 20 as missing and excluded all loci with fewer than five genotyped individuals per species and all loci where the rarer allele was observed <3 times across all individuals.

A particular challenge in *de novo* assembly is the correct distinction between orthologs and paralogs. Specifically, the goal is to distinguish between true variation at a single locus and artefactual variation created by merging nonorthologous reads into the same locus (e.g. Stapely *et al.* 2010). If two paralogous loci fixed for alternative variants are wrongly combined, we expect to observe a pronounced excess of heterozygous individuals (e.g. Hohenlohe *et al.* 2011). Consequently, we excluded loci from subsequent analyses if the observed heterozygosity in one or more species was higher than 0.5, the maximum Hardy–Weinberg heterozygosity at a biallelic locus. All filtering steps were carried out with custom-made python scripts. Estimates of genetic diversity in each species are given in Table S2 (Supporting information).

To assess the effect of specific parameters on the inferred number of loci and SNPs, we re-ran this pipeline twice with (i) fewer and (ii) more mismatches tolerated within loci in the *de novo* assembly and the subsequent mapping of the reads to the consensus sequences from the assembly. Please refer to the Supporting information for a detailed description of these analyses and a discussion of the results. Briefly, we found that as more mismatches are tolerated within a RAD locus, more reads are used in the assembly, and they are merged into fewer loci (Table S1, Supporting information). This increases the number of polymorphic sites overall, as well as the proportion of RAD loci containing more than one SNP. It is encouraging to find that analyses based on the three different assemblies produce highly consistent results. For example, we detected very similar outlier proportions (Fig. S1, Supporting information), and F_{ST} estimates between all species pairs were also highly correlated across all three assemblies (correlation coefficients $\geq$ 0.98 in all cases).

Population genomic analysis

The quality filtered vcf files were converted into the file formats necessary for subsequent analyses using PGDSPIDER version 2.0.1.5 (Lischer & Excoffier 2012).

To identify SNPs showing evidence of divergent selection between species, we carried out outlier scans between all species pairs in BAYESCAN version 2.1 using default settings (Foll & Gaggiotti 2008). In particular, the prior odds were set to 10 corresponding to a prior belief that the neutral model is 10 times more likely than the model with selection at any given locus. A locus was considered to be an outlier if the posterior odds favouring a model with selection were above a threshold set by the software to ensure a false discovery rate of $\leq$20%. To evaluate the effect of the prior, all pairwise outlier scans were repeated with prior odds 1 corresponding to equal prior probabilities for the models with and without a selection term. All SNPs significant in this last analysis were removed from the data sets to obtain estimates of neutral F_{ST} between all species pairs in ARLEQUIN 3.5.1.2 (Excoffier *et al.* 2005). The significance of the F_{ST} estimates was assessed based on 10 000 permutations.

We then investigated the genetic substructure in our data with STRUCTURE v. 2.3.2.1 (Pritchard *et al.* 2000; Falush *et al.* 2003). We used a burn-in of 500K steps followed by another 500K MCMC steps, assumed an admixture model and correlated allele frequencies, and included no prior information on taxon identity. A first analysis was based on the full data set of 10 663 SNPs genotyped in at least five individuals per species. We varied the number of groups (K) from 1 to 8 with 10 independent runs for each value of K. We then ordered all SNPs based on the locus-specific F_{ST} as calculated in ARLEQUIN across all five species ('global' F_{ST}, from lowest to highest), and created data subsets containing different slices from this distribution: intermediate global F_{ST} (all SNPs between 25th and 75th percentile of the F_{ST} distribution corresponding to F_{ST} = 0.01–0.13; in the following, this will be expressed as 25–75%), high global F_{ST} (75–99%, F_{ST} = 0.13–0.51) and top global F_{ST} (99–100%, F_{ST} = 0.51–1.0). We used the same run parameters as in the full analysis but investigated only values of K = 1–5. An additional six data subsets were investigated from K = 1–5 with only 3 runs per K (0–50%, 0–75%, 0–90%, 0–99%, 25–35% and 25–50%). The support for different values of K was assessed from the likelihood distribution and visual inspection of the STRUCTURE barplots. In particular, the consistency of the inferred clusters across runs was an important criterion to select the optimal K. Results were averaged across replicates using CLUMPP 1.1.2 (Jakobsson & Rosenberg 2007) and graphically displayed with DISTRUCT 1.1 (Rosenberg 2004).

Pairwise F_{ST} between all species pairs was calculated for the different data subsets in ARLEQUIN with significance assessed based on 10 000 permutations. We investigated the correlation between pairwise F_{ST}s calculated from different data subsets using Mantel tests in FSTAT. Finally, analyses of molecular variance (AMOVA; Excoffier *et al.* 1992) were performed on selected data sets with populations grouped either by genus or by male colouration. The analyses were repeated without *P.* sp. 'pink anal fin' that is genetically closer to *M. mbipi* than to *Pundamilia* (see below). The significance of the variance components was assessed based on 1000 permutations.

Finally, the full data set and the three subsets (intermediate, high and top) were used to infer phylogenies using a maximum-likelihood approach in RAxML (Stamatakis 2006). We used a GTR+gamma model of sequence evolution, as recommended and justified by the authors of the program in the version 7.0.4 manual, and to account for uncertainty in the estimation of the topology, we used RAxML's rapid bootstrap algorithm with 100 bootstrap replicates for the full data and 10 000 bootstrap replicates for each data subset (Stamatakis *et al.* 2008), along with a single, full maximum-likelihood tree search.

Results

Loci showing elevated divergence between species pairs

The average proportion of outliers detected by BAYESCAN was 0.71%, ranging from a minimum of 0.42% in the *M. mbipi* vs. *P.* sp. 'pink anal fin' comparison to a maximum of 0.91% between *P. pundamilia* and *P. nyererei*. At two-thirds of the outlier SNPs (65%), outlier behaviour was observed only in a single comparison between species. The vast majority (92%) of cases where a SNP was a significant outlier in multiple pairwise comparisons repeatedly involved the same species. For all remaining SNPs, we investigated how many were detected as significant outliers in multiple independent comparisons between genera and/or in multiple independent comparisons between male colour types. We found eight such replicated between-colour outliers (blue + turquoise in Fig. 1) and eight between-genus outliers (green + turquoise in Fig. 1).

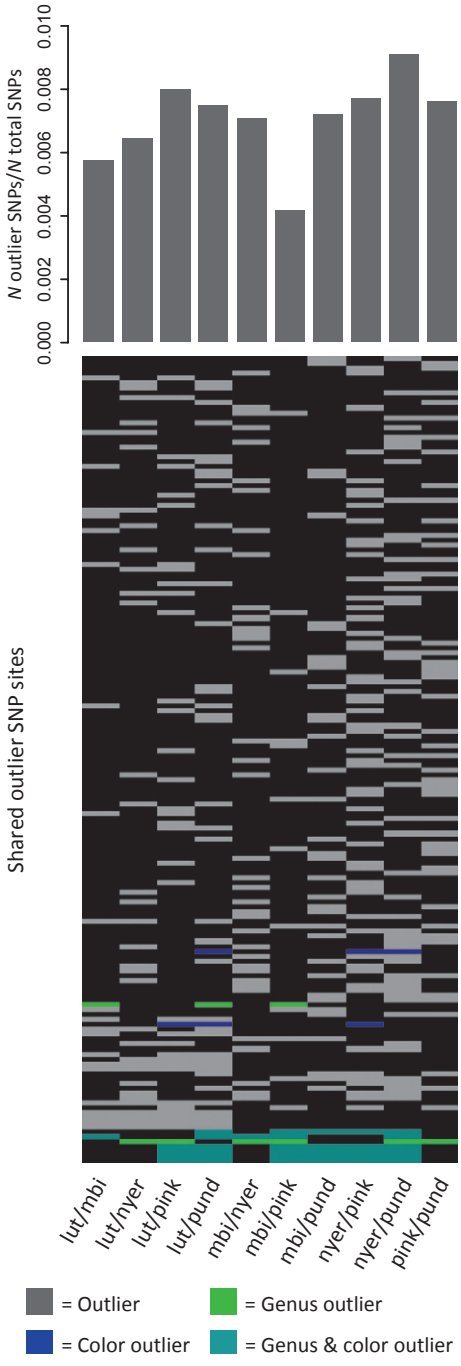

Fig. 1 Results of outlier scans for all pairwise comparisons among the five cichlid species. The barplot indicates the proportion of SNPs detected as significant outliers in each comparison. In the bottom panel, each column represents a pairwise comparison and each row a SNP site showing outlier behaviour in ≥ 2 comparisons. We were specifically interested in identifying SNPs detected as outliers in ≥ 2 independent comparisons between the two genera and/or the two colour types. If this criterion was satisfied for genus and/or colour, we coloured all significant comparisons at that locus. Green = between-genus outlier in ≥ 2 independent comparisons; blue = between-colour-type outlier in ≥ 2 independent comparisons; turquoise = outlier in ≥ 2 of both between-genus and between-colour comparisons. All other significant comparisons are indicated in grey. SNPs are ordered from top to bottom by the number of comparisons in which they were detected as outliers among the ten pairwise comparisons. lut = *Mbipia lutea*; mbi = *M. mbipi*; nyer = *Pundamilia nyererei*; pink = *P.* sp. 'pink anal fin'; pund = *P. pundamilia*.

The results from the analysis assuming even prior probabilities for the models with and without selection (prior odds 1) were qualitatively similar (correlation coefficient 0.90), but the portion of outliers was, on average, 7.5 times higher with even prior odds. In this analysis, we identified 62 replicated between-colour outliers and 85 replicated between-genus outliers (Fig. S2, Supporting information).

Patterns of genomic differentiation

Genetic substructure was observed in the STRUCTURE analyses of the full data set and of all subsets containing some markers from the upper half of the F_{ST} distribution (0–75%, 0–90%, 0–99%, 25–75%, 75–99% and 99–100%). Very little substructure was observed in subsets containing only markers from the lower half of the distribution (25–35% and 25–50%). The full data set and three non-overlapping subsets were investigated in more detail. In the following, we will refer to these subsets as intermediate (25–75%), high (75–99%) and top (99–100%) to indicate the position of a given 'slice' in the distribution of all locus-specific F_{ST} values ordered from smallest to largest (Fig. 2a). In all four cases, the likelihood increased substantially up to $K = 4$ with a smaller but discernible further increase between $K = 4$ and $K = 5$ (Fig. S4, Supporting information). In many analyses, the inferred clusters varied between runs with up to five different solutions found for a given data set and value of K. Here, we consider a value of K to be optimal if it has both a high likelihood and produces consistent results across all 10 replicate STRUCTURE runs. Under this criterion, four genetic clusters were inferred based on the intermediate SNPs and five in the remaining two data subsets (Fig. 2c) and the full data set (Fig. S3a, Supporting information).

The F_{ST} values between species pairs were significantly correlated in all six possible pairwise comparisons between the four data (sub)sets ($R^2 \geq 0.64$, Mantel $P < 0.01$ in all cases). However, we observed striking differences between the four data sets in the relationships among the five species as inferred by STRUCTURE and the phylogenetic analyses, with the following two main results:

1 The nearer to the upper tail of the F_{ST} distribution our SNP sets were, the more pronounced was the separation inferred by STRUCTURE between the five clusters corresponding to the five phenotypically defined species: the intermediate SNPs assigned all *M. mbipi* and *P.* sp. 'pink anal fin' individuals to the same genetic cluster and showed consistent evidence of contributions of the *M. lutea*-like genetic cluster (light orange in Fig. 2c) to *M. mbipi*. Hence, this result is consistent with past hybridization between two sets of taxa: *M. mbipi* and *P.* sp. 'pink anal fin', and *M. mbipi* and *M. lutea*. This evidence for introgression became weaker among the high SNPs and completely disappeared in the top 1% (Fig. 2c).

2 The intermediate and high data sets mostly revealed the strongest structure between two species groups: *P. pundamilia* and *P. nyererei*, and the triplet *M. mbipi*, *M. lutea* and *P.* sp. 'pink anal fin'. These groups reflect the morphology-based genus assignment with the exception of *P.* sp. 'pink anal fin'. We observed this grouping seven (for intermediate SNPs) and five times (for high SNPs) among each of the 10 STRUCTURE replicates, and it was consistently better supported than a grouping based on male colouration, which was observed only once or three times, respectively (Fig. 2b). The ML tree based on the intermediate SNPs was poorly resolved (Fig. 3a). The phylogenetic analysis of the high data subset recovered the same two 'nearly generic' clusters observed in the STRUCTURE results-*P. pundamilia*/*P. nyererei* and *M. mbipi*/*M. lutea*/*P.* sp. 'pink anal fin'-with high bootstrap support (Fig. 3b).

A different pattern was observed in the very upper tail of the F_{ST} distribution (top SNPs). Here, the first genetic split inferred by STRUCTURE (i.e. at $K = 2$) was clearly between species of different male nuptial colouration (Fig. 2b; 9 of 10 replicates), a split that was also supported in phylogenetic analyses (Fig. 3c).

These findings were also supported by the results of the analyses of molecular variance based on the three non-overlapping data subsets (Table 2). The proportion of the genetic variance explained by genus was not significant when the analysis was based on all five species (Table 2a, top). However, this component became significant for the intermediate and high subsets of SNPs (but not the top SNPs) when *P.* sp. 'pink anal fin' was excluded (Table 2b, top). This exclusion seems justified based on our results here which suggest hybridization between *Pundamilia* and *Mbipia* at the origin of this species and that it is in fact genetically closer to *Mbipia mbipi* than to the two other *Pundamilia* species. The proportion of the variance explained by differences between colour types showed a pattern opposite to that observed for the between-genus component: the between-colour component was considerably higher in the analysis based on the top SNPs than in the other two data subsets irrespective of whether *P.* sp. 'pink anal fin' was included or not (Table 2, bottom).

Discussion

Using 10 000 SNPs from 7000 loci generated by *de novo* assembly of sequenced RAD tags, we find

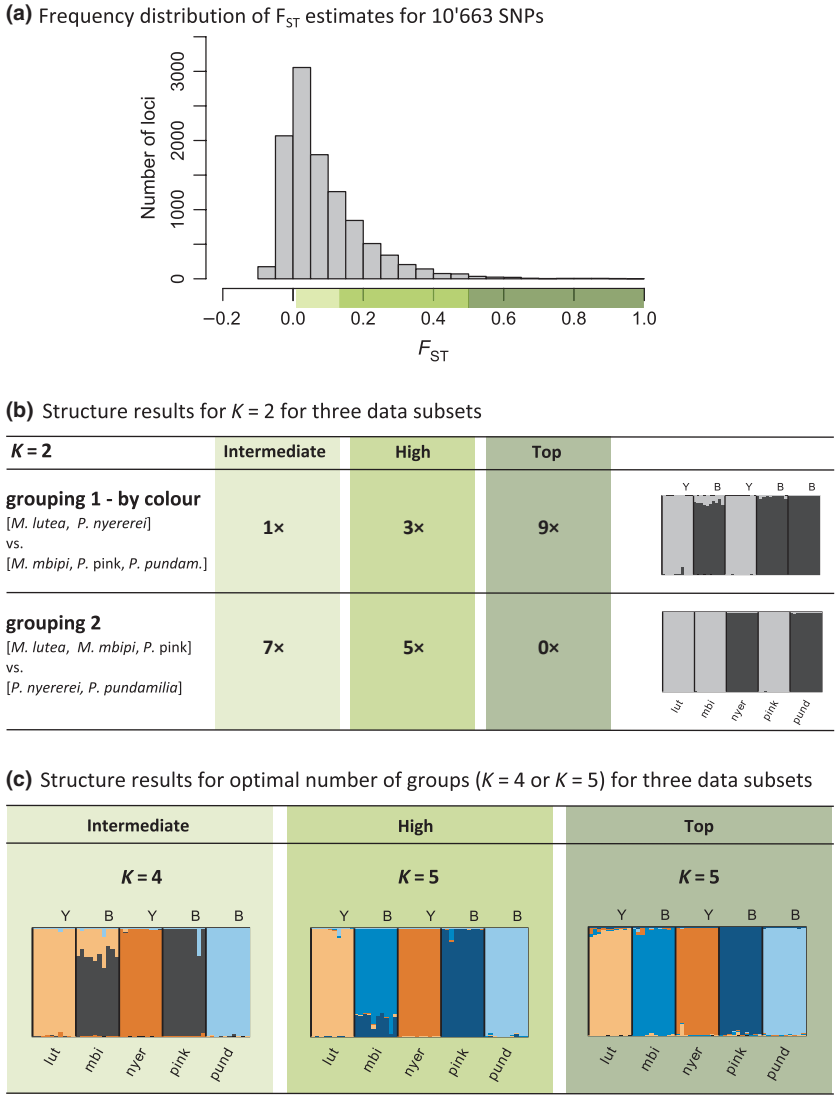

Fig. 2 Population structure inferred from different subsets of loci. a) Frequency distribution of F_{ST} values calculated across all five species (global F_{ST}) at 10 663 SNPs. All SNPs were then arranged in order of increasing F_{ST}, and different slices from this distribution were used for STRUCTURE analyses. The approximate range of F_{ST} values included in each slice is indicated by the green bars below panel a). (b) STRUCTURE results for $K = 2$. In most cases, several alternative solutions were observed. Here, we present only the two dominant solutions and report the number of times each was observed among a total of 10 replicate runs in each data subset. Grouping 1: the species cluster according to male nuptial colouration as indicated by the letters above the STRUCTURE barplot (Y = yellow; B = blue). Grouping 2: species group by genera with the exception of *P.* sp. 'pink anal fin' that clusters with the two *Mbipia* species. Each clustering solution is illustrated by a STRUCTURE barplot based on the high data subset (see Fig. S4 for all other plots), with results averaged across all runs supporting this particular grouping. (c) Results of the STRUCTURE analysis for the number of clusters best supported in each data set ($K = 4$ or $K = 5$, as indicated), with the results averaged across 10 runs. Intermediate = 5331 SNPs with F_{ST} values between the 25th and 75th percentiles of all locus-specific F_{ST} values ordered from lowest to highest; high = 75th–99th percentile, 2559 SNPs; top = above 99th percentile, 107 SNPs. lut = *Mbipia lutea*; mbi = *M. mbipi*; nyer = *Pundamilia nyererei*; pink = *P.* sp. 'pink anal fin'; pund = *P. pundamilia*.

clear evidence of distinct, non-overlapping genetic groups corresponding to five phenotypically defined species of Lake Victoria cichlids sampled in full sympatry at one island. This is striking, because previous studies based on DNA sequences failed to differentiate species altogether (e.g. Nagl *et al.* 1998; Samonte *et al.* 2007), and studies using polymorphisms in microsatellites or AFLP restriction sites, while finding that species were differentiated, failed to assign individuals reliably (Seehausen *et al.* 2008;

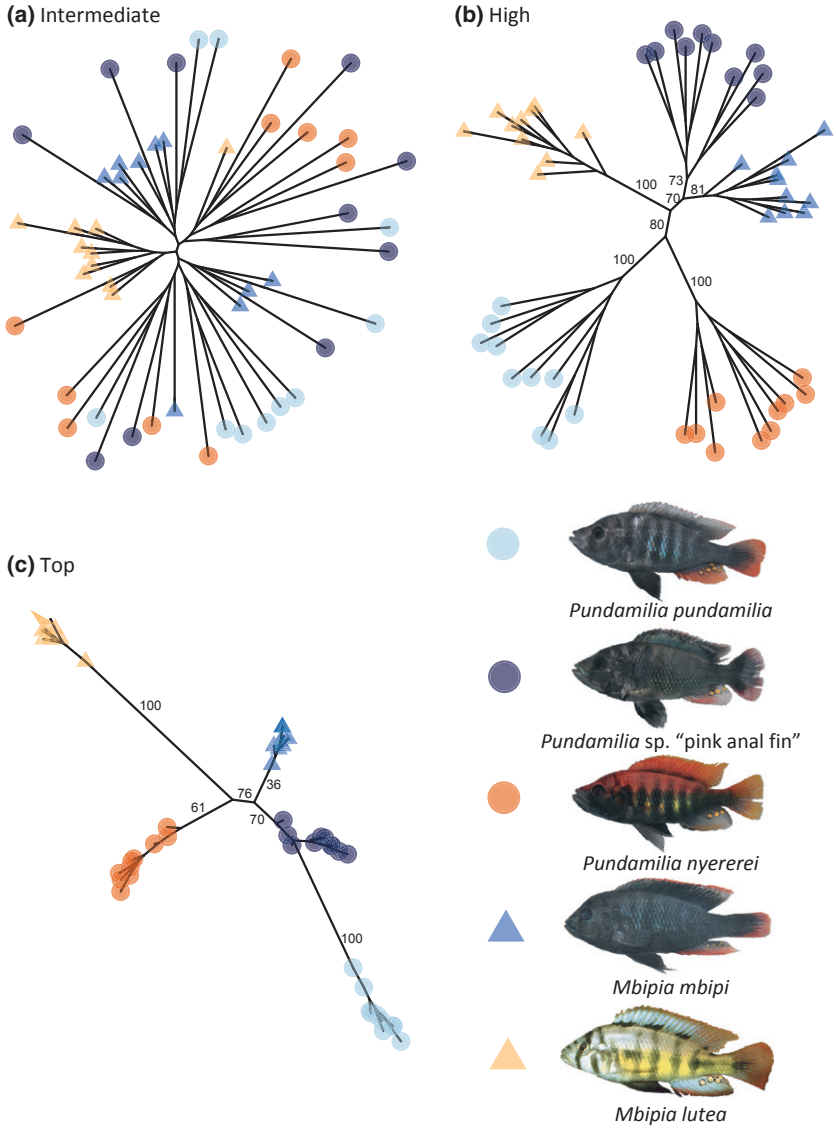

Fig. 3 Phylogenetic trees based on different data subsets: (a) Intermediate: 5331 SNPs with global F_{ST} values between the 25th and 75th percentiles of a distribution of all locus-specific estimates ordered from lowest to highest; (b) High: 75th–99th percentile, 2559 SNPs; (c) Top: above 99th percentile, 107 SNPs. Topologies shown are the best tree from a full maximum-likelihood search. Tip colours represent the species. The colours are consistent with male nuptial colouration and match Fig. 2c. Triangles indicate *Mbipia* spp. and circles *Pundamilia* spp. Bootstrap support values are based on 10 000 rounds of bootstrapping using RAxML's rapid bootstrapping algorithm, and are shown only if $\geq 50\%$ and are not shown within species groups.

Magalhaes *et al.* 2009, 2012; Konijnendijk *et al.* 2011). This highlights the power of large SNP data sets as generated with genotyping-by-sequencing approaches to resolve the population genetic and phylogenetic structure of even the most rapid and difficult to resolve sympatric radiations, even in the absence of a reference genome (see also Wagner *et al.* 2012). Furthermore, we find that the relationships estimated from the loci that best differentiate species do not follow the phylogenetic relationships observed in analyses using other portions of the genome.

Loci showing elevated divergence between species pairs

In our analyses, many SNPs show outlier behaviour in only a single comparison. Another very large proportion is significant in multiple comparisons, but these repeatedly involve the same species. This pattern suggests unusual allele frequencies in one species, possibly because of species-specific selection pressures. Similar observations have been reported in intraspecific outlier scans where multiple populations were compared (e.g. Akey *et al.* 2002; Williams & Oleksiak 2008; Keller *et al.*

Table 2 Percent of the variation explained by different groupings based on analyses of molecular variance of three data subsets with species grouped by genus (top) or male nuptial colouration (bottom)

	Intermediate	High	Top
Grouped by genus			
(a) [*P. pundamilia, P. nyererei, P.* sp. 'pink anal fin'] [*M. mbipi, M. lutea*]			
Between genera	0.16	0.32	0.65
Among species within genera	6.15**	24.11**	65.24**
(b) [*P. pundamilia, P. nyererei*] [*M. mbipi, M. lutea*]			
Between genera	0.58**	2.43**	−5.94
Among species within genera	6.13**	23.74**	72.04**
Grouped by colour			
(a) [*P. pundamilia, P.* sp. 'pink anal fin', *M. mbipi*] [*M. lutea, P. nyererei*]			
Between colours	0.42**	2.15**	14.41**
Among species within colours	5.99**	22.83**	53.38**
(b) [*P. pundamilia, M. mbipi*] [*M. lutea, P. nyererei*]			
Between colours	0.05	0.03	8.67*
Among species within colours	6.49**	25.56**	59.06**

(a) All five species. (b) Excluding *Pundamilia* sp. 'pink anal fin'. *$P < 0.001$; **$P < 0.0001$.

2011; E. Bezault, G. Dheyongera, S. Mwaiko, I. S. Magalhaes, O. Seehausen, Unpublished data).

Still, we also find a number of SNPs that show outlier behaviour, consistent with the action of divergent selection, in at least two fully independent comparisons between genera (green + turquoise in Fig. 1) or between colour types (blue + turquoise in Fig. 1). Such parallel patterns are predicted for body colouration where the underlying genes are suspected to be conserved across these cichlids, and alleles may have been transferred between lineages by occasional hybridization (Seehausen *et al.* 1999; Magalhaes & Seehausen 2010; see also discussion below). Parallel patterns at the gene level have previously been shown for opsin genes that determine light absorbance properties of the retina. Here, sister species from different light environments are often characterized by distinct divergent alleles, which are at the same time often shared between distantly related species (Seehausen *et al.* 2008).

It must be kept in mind that the loci detected as repeated outliers here almost certainly represent only the most strongly differentiated regions of the genome given the low power of pairwise outlier tests (Fischer *et al.* 2011). However, it is encouraging that it is possible to identify candidate loci potentially involved in adaptive divergence between these genera or replicate adaptive divergence between colour types. These results indicate that either many loci are associated with these traits in Lake Victoria cichlid genomes or the signal of divergent selection extends across few but large linkage blocks. Such an interpretation is suggested by the fact that the detection of candidate loci requires a combination of several fortunate circumstances: an *SbfI* restriction site has to be present near the target of selection, polymorphic sites have to exist within 80 bp either side of the restriction site, and the read depth has to be sufficient for the locus to be genotyped in multiple individuals per species. Hence, a substantial proportion of the loci showing strong evidence of adaptive divergence are likely to be missed. For example, *bmp4*, a gene involved in tooth development (e.g. Albertson & Kocher 2006), should be a prime candidate locus underlying the divergence between the unicuspid teeth of *Pundamilia* and the bicuspid teeth of *Mbipia* species. The *bmp4* coding sequence (e.g. GenBank accession number AB084660) indeed contains an *SbfI* restriction site, but our reads did not include any polymorphic sites at this RAD locus, precluding an analysis of the patterns of genetic differentiation for these species using the RAD approach.

The total number of outliers detected ranged from a few dozen to several hundred per species pair depending on the analysis. While the chromosomal location of these loci is currently unknown, it is likely that multiple independent regions are involved in the divergence between each species pair. A single region containing even the lowest observed number of outliers (41) could be as large as 1.2 Mb if we assume one *SbfI* site every 30 Kb, the approximate average spacing of restriction sites expected from the predicted number of *c.* 30 K *SbfI* sites and a genome size of *c.* 950 Mb (Sanetra *et al.* 2009). While gene flow can be reduced across regions of several megabases as a result of divergence hitchhiking (Via 2012), signals of selection are likely to be restricted to much shorter chromosomal regions if adaptation is from standing genetic variation (Barrett & Schluter 2008; Counterman *et al.* 2010) as is likely often the case in these cichlids. For example, a comparison of cichlids from different light environments found that genetic differentiation was elevated within a window of <10 kb around a target of divergent selection (long wavelength-sensitive opsin gene; Terai *et al.* 2006).

Patterns of genomic differentiation

The comparison of both the STRUCTURE results and the phylogenetic trees inferred from different subsets of the data showed that evidence of genetic subdivision between species is widespread across the genome. More

than half of the SNPs in our data set show some differentiation between species as indicated, for example, by the pairwise F_{ST} estimates obtained from the intermediate data set (which contains the SNPs between the lower and upper quartiles of the F_{ST} distribution). This section of the data should produce good estimates of the average genomic divergence between species as it omits loci from the tails of the distribution, some of which are expected to be under balancing or divergent selection. The obtained F_{ST} estimates (Table S3, Supporting information) range between 0.025 between *M. mbipi* and *P.* sp. 'pink anal fin' and 0.061 between *M. lutea* and *P. pundamilia*. F_{ST} between *P. pundamilia* and *P. nyererei* at Makobe Island has previously been estimated at 0.026 from microsatellites (Seehausen *et al.* 2008) and 0.031 based on non-outlier AFLP markers (E. Bezault, G. Dheyongera, S. Mwaiko, I. S. Magalhaes, O. Seehausen, Unpublished data), compared with an estimate of 0.055 obtained here. Hence, differentiation as measured from our intermediate SNP data set is broadly comparable with estimates from classic 'neutral' markers.

The genetic relationships observed among the species change as we move up in the F_{ST} distribution. We interpret this change as a progressive erosion of the signal of neutral phylogenetic history as we move towards the more highly differentiated SNPs. The presumably neutral structure inferred from the intermediate and high data subsets suggests that the two genera are not reciprocally monophyletic. In these data subsets, the first split detected by STRUCTURE (at $K = 2$) does not fully coincide with the two genera, and results from the analyses of molecular variance also show that genus does not explain a significant proportion of the total variance. Both types of analyses demonstrate an unexpectedly close genetic relationship between *P.* sp. 'pink anal fin' and *M. mbipi*, which is also consistent with the phylogenetic tree based on the high SNPs (Fig. 3b), a result at odds with rich phenotypic data that groups *P.* sp. 'pink anal fin' with other *Pundamilia* species.

After excluding *P.* sp. 'pink anal fin', the genetic variation explained by the between-genus component was small but significant in the intermediate and high data subsets (AMOVA; Table 2). Similarly, an AFLP-based analysis with much more extensive taxon sampling found that a small but significant proportion of the total genetic variance was associated with genera in Lake Victoria cichlids (Bezault *et al.* 2011). These findings suggest that the average genomic divergence of species tends to be slightly lower within than between genera, as would be expected if these genera are indeed meaningful groups that reflect evolutionary relationships.

Importantly, very different genetic relationships among the species are inferred based on the 107 SNPs with the highest global F_{ST} estimates, suggesting that the loci experiencing the strongest divergent selection between species do not obey historical evolutionary relationships. Most prominently, these top 1% of SNPs are enriched for markers that group species by colour types and markers that differentiate all five species. These patterns are illustrated by the contrasting relationships among species in the phylogenetic trees inferred from the high and top subsets of the data (Fig. 3b,c) and by the clear split into five distinct species clusters only in the STRUCTURE results based on the 107 top SNPs (Fig. 2c).

The high resolution of species boundaries obtained from these 107 SNPs is not necessarily expected as these markers were selected based on high global F_{ST} across all five species, which could, in principle, result from unusual allele frequencies in a single species (e.g. Shriver *et al.* 2004). It suggests that loci important in speciation and/or the maintenance of species boundaries are enriched in this data subset. In other words, extreme allele frequency differences are observed between different species or groups of species at different SNPs, which suggests that divergent selection played an important role in speciation.

The observed grouping by colour type supported by several of our results based on the top 1% SNPs suggests that this data subset includes at least some loci where non-sister species of the same colour are genetically more similar to each other than they are to their differently coloured sister species. We identified 13 loci that consistently show contrasting allele frequencies between all three species with blue male colouration and the two species with yellow male colouration (allele frequency difference between colour types >0.25; data not shown). Male colouration is known to play an important role in speciation and reproductive isolation among Lake Victoria cichlid fishes (Seehausen & Van Alphen 1999; Seehausen & Schluter 2004). In particular, speciation involves parallel divergence along the same yellow–blue axis of male colouration in a number of genera including those studied here (Seehausen & Van Alphen 1999; Seehausen *et al.* 1999, 2008). One idea to explain this pattern is that such repeated bouts of disruptive selection on what are probably the same genes could have been made possible by the specific genetic architecture of this trait, preventing the complete erosion of genetic variation in the course of selective sweeps (Seehausen *et al.* 1999; Magalhaes & Seehausen 2010). Alternatively or additionally, allelic variation may have been reintroduced through occasional hybridization events (Seehausen 2004; *Heliconius* Genome Consortium 2012; Jones *et al.* 2012). Our finding that the loci that best differentiate these five species do not follow the phylogenetic rela-

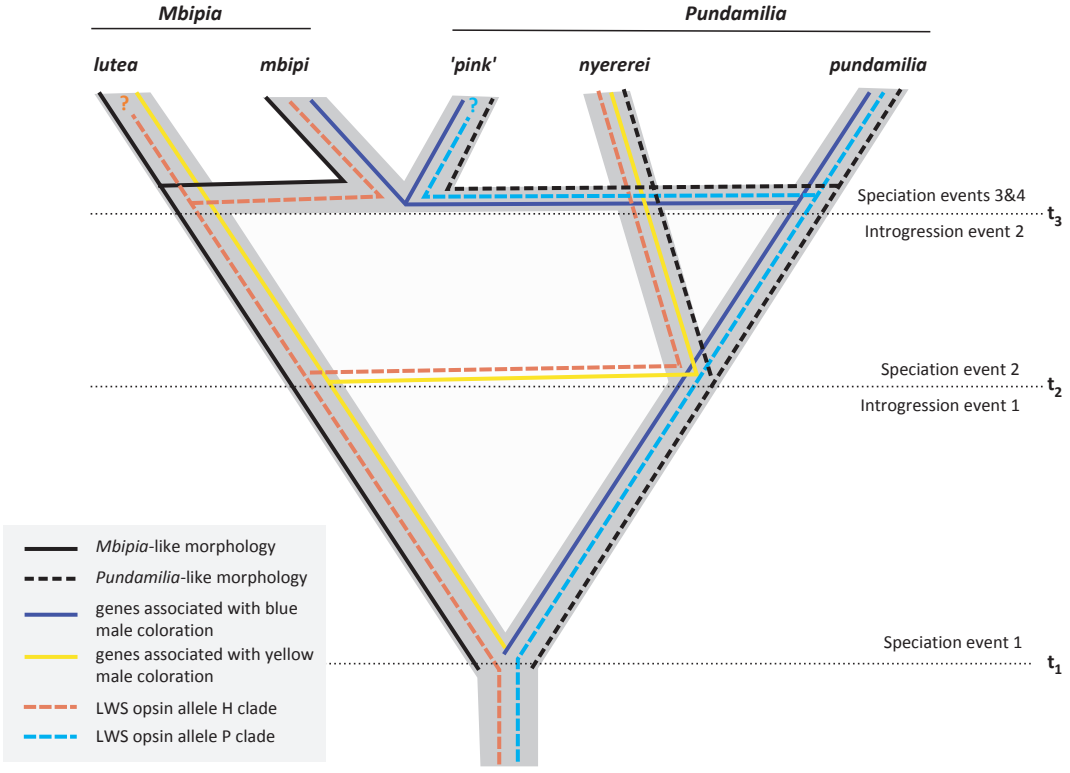

Fig. 4 A scenario for the evolution of the five species that is consistent with morphological and molecular data. Time point 1 (t_1): first speciation event between *Mbipia lutea*-like and *Pundamilia pundamilia*-like ancestor. t_2: secondary hybridization leads to introgression of allelic variation (genes associated with yellow colouration; LWS opsin H clade) into *Pundamilia*, thus facilitating speciation event 2 within *Pundamilia*. t_3: a second intergeneric hybridization event precedes speciation events 3 and 4 that give rise to *M. mbipi* and *P.* sp. 'pink anal fin'. LWS opsin data are not available for two species. In these cases, the most likely allele clade is shown together with a question mark.

tionships estimated from the majority of the genomic loci suggests that divergent selection during some of the speciation events may indeed have recruited variation introduced by hybridization among non-sister species.

Reconciling the trees: a scenario for hybrid speciation

Attempting to develop a scenario that reconciles the phylogenetic and population genetic analyses presented here, we can envision the following evolutionary scenario (Fig. 4): *P. pundamilia* and *M. lutea* are the genetically most divergent among our five study species, and this is consistent in all our SNP subsets. They are also certainly phenotypically most distinct. *Pundamilia nyererei* is the most distinct among the remaining three species, both in terms of population genetic and phylogenetic analyses, yet its phylogenetic relationships when estimated from the top slice of the F_{ST} distribution (closely related to *M. lutea*) vs. the rest of the genome (closely related to *P. pundamilia*) are inconsistent. We propose that the most parsimonious scenario consistent with all of these results involves a first split (at t_1; Fig. 4) between a yellow *Mbipia* ancestor and a blue *Pundamilia* ancestor. This speciation event was followed by some hybridization between the two taxa (at t_2; Fig. 4), which lead to the introgression of allelic variation from *Mbipia* into *Pundamilia*, followed by divergent selection between colouration genes within *Pundamilia* and, ultimately, speciation into *P. pundamilia* and *P. nyererei*. *Pundamilia nyererei* exhibits some traits potentially derived through introgression from *Mbipia* (i.e. colour). This scenario is also consistent with patterns of allele sharing observed at the *LWS* opsin gene in earlier work (Fig. 4): while *P. pundamilia* and *P. nyererei* at Makobe are fixed for alleles from divergent opsin clades, *P. nyererei* shares its alleles with *Mbipia mbipi*, and they are embedded in a clade with several other alleles that occur in *Mbipia* but not in *P. nyererei* (note that we do not have *LWS* data for *M. lutea*, but it is likely that they share alleles from the same allele clade that is widespread in shallow water algae scrapers; Terai *et al.* 2006; Seehausen *et al.* 2008).

Mbipia mbipi and *P.* sp. 'pink anal fin' are the least well differentiated of the five species studied here. They are poorly differentiated from each other across most of their genome (Fig. 2c), they show the fewest pairwise outlier loci (Fig. 1), and they are the least resolved species groups in the tree estimated from the top SNP data subset (Fig. 3c). We observe low support for the monophyly of each species and only very short branches in the top SNP data (Fig. 3c), suggesting most of their alleles at these SNPs are shared with other species. To verify whether this is indeed the case, we examined the allele frequencies at all 107 top SNPs and identified all cases where the observed allele frequencies in one species were very different from those in the remaining four species (frequency difference between the two groups ≥ 0.6). This condition would be met, for example, at a SNP site where nucleotide A has the following frequencies in each of five species: 0%, 12%, 38%, 2% and 100%. In the following, we would consider this locus to have a 'unique allele frequency' in the fifth species, and we expect that species of hybrid ancestry would have fewer such unique allele frequencies than genetically very distinct species.

We found that *M. lutea* and *P. pundamilia* had unique allele frequencies at a much larger number of loci (23 and 17, respectively) than the other three species (Fig. 5a). *Mbipia mbipi* had just three such loci, *P.* sp. 'pink anal fin' just one, while *P. nyererei* was intermediate with eight uniquely differentiated loci. Taken together, these observations imply that both *M. mbipi* and *P.* sp. 'pink anal fin' share a history of admixture between *Mbipia* and *Pundamilia*. This hybridization episode would have been more recent than that inferred in the ancestry of *P. nyererei*, and it seems possible that both species derive from the same episode of hybridization between a population of *M. lutea* and one of *P. pundamilia* (at t_3 in Fig. 4).

Importantly, the allele frequencies observed in *M. mbipi* and *P.* sp. 'pink anal fin' are not consistently intermediate between those of *M. lutea* and *P. pundamilia* at individual loci. Instead, the two species carry *M. lutea* alleles at some loci and *P. pundamilia* alleles at others. To illustrate this pattern, we identified all top SNPs where the frequency of a particular allele was similar in a hybrid species and one parental species, but showed a frequency difference of ≥ 0.6, that is, a unique allele frequency, in the second putative parent (as illustrated for *M. mbipi* in Fig. 5b). Using this criterion, of the total 107 top SNPs, we identified 27 where the allele frequency in *M. mbipi* was similar to *M. lutea* and 32 where it was similar to *P. pundamilia*. In *P.* sp. 'pink anal fin', we identified 21 SNPs where the allele frequencies were *M. lutea*-like and 40 where they were *P. pundamilia*-like. These results suggest significant contributions from both putative parental species to the hybrid genomes.

Based on the current analyses, we cannot confidently exclude the possibility that incomplete lineage sorting (ILS) of ancestral variation has contributed to the inconsistencies among the genealogical relationships inferred from different sets of loci. However, it seems difficult to envisage how ILS alone could produce the observed pattern of highly admixed genomes in some species but not in others. Coalescent-based simulations will provide future opportunities for explicit tests of the relative roles of ILS, hybridization and selection in the origin and spread of alleles in these populations.

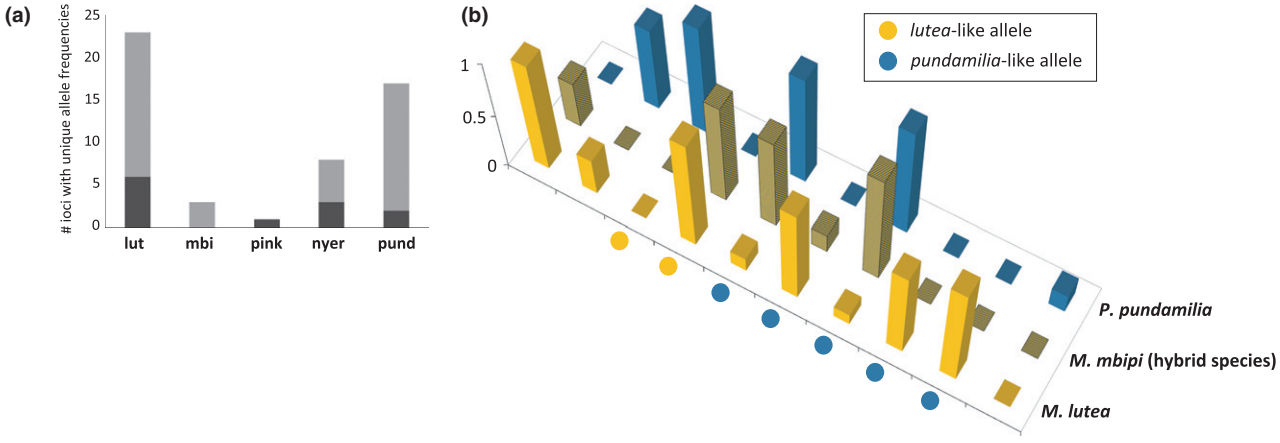

Fig. 5 (a) Number of loci with unique allele frequencies in each species. An allele is considered to be unique in a species if its frequency differs by ≥ 0.6 (dark and light grey) or ≥ 0.8 (dark grey only) from the frequencies observed in the remaining four species. (b) Frequency of one allele at 10 SNP sites from among the 107 SNPs with the highest global F_{ST} (top SNPs) in *Mbipia lutea*, *Mbipia mbipi* and *Pundamilia pundamilia*. Yellow circles indicate sites where the allele frequency in *M. mbipi* is similar to that in *M. lutea* and blue circles SNPs where *M. mbipi* is similar to *P. pundamilia*.

Conclusions

For the first time in Lake Victoria cichlids, a genome-scale data set derived from genotyping-by-sequencing approaches allowed us to contrast phylogenetic relationships reconstructed from distinct subsets of markers characterized to different extents by selection history, introgression and ancestry. The rapid species radiation of Lake Victoria cichlid fishes is associated with signatures of repeated hybridization. Our results suggest that hybridization lead to reshuffling of gene complexes associated with adaptation and mate choice, resulting in hybrid species with novel combinations of ecology and mating behaviour, not unlike the case of *Heliconius* butterflies (*Heliconius* Genome Consortium 2012). These results suggest an important contribution of episodes of hybridization to this extraordinarily fast adaptive radiation of cichlid fishes.

Acknowledgements

We are extremely grateful to Heidi Lischer and Stefan Zoller for their invaluable help with the analyses. Computational resources were provided by the Genetic Diversity Centre, ETH Zürich, Switzerland and the Computational Biology Service Unit at Cornell University, which is partially funded by Microsoft Corporation. We would like to thank Rahel Thommen for assistance with laboratory work, Mhoja Kayeba, Mohammed Haluna, Martine Maan, Erwin Ripmeester and Dora Selz-Affolter for assistance with sample collection, and Joana Meier, David Marques, Matthieu Foll and Julian Catchen for helpful discussion. Thanks also to the Tanzania Commission for Science and Technology (COSTECH) for providing permits to collect samples in Lake Victoria, and to the Tanzanian Fisheries Research Institute (Y.L. Budeba, B.P. Ngatunga, E.F.B. Katunzi and H.D.J. Mrosso) for hospitality and facilities. We are grateful to the editor and three anonymous reviewers for their constructive comments on an earlier version of this manuscript. The work was supported through Swiss National Science Foundation grant 31003A-118293 to O. Seehausen.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Albertson RC, Kocher TD (2006) Genetic and developmental basis of cichlid trophic diversity. *Heredity*, **97**, 211–221.

Allender CJ, Seehausen O, Knight ME, Turner GF, Maclean N (2003) Divergent selection during speciation of Lake Malawi cichlid fishes inferred from parallel radiations in nuptial coloration. *Proceedings of the National Academy of Sciences*, **100**, 14074–14079.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**, e3376.

Barchi L, Lanteri S, Portis E *et al.* (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics*, **12**, 304.

Barrett RDH, Schluter D (2008) Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, **23**, 38–44.

Beaumont MA (2005) Adaptation and speciation: what can F_{st} tell us? *Trends in Ecology & Evolution*, **20**, 435–440.

Bezault E, Mwaiko S, Seehausen O (2011) Population genomic tests of models of adaptive radiation in Lake Victoria region cichlid fish. *Evolution*, **65**, 3318–3397.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 171–182.

Counterman BA, Araujo-Perez F, Hines HM *et al.* (2010) Genomic hotspots for adaptation: the population genetics of Müllerian mimicry in *Heliconius erato*. *PLoS Genetics*, **6**, e1000796.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Egan SP, Nosil P, Funk DJ (2008) Selection and genomic differentiation during ecological speciation: isolating the contributions of host association via a comparative genome scan of *Neochlamisus bebbinae* leaf beetles. *Evolution*, **62**, 1162–1181.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In: *Molecular Methods for Evolutionary Genetics* (eds Orgogonzo V, Rockman MV), pp. 157–178. Humana Press, NY.

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, **1**, 47–50.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479–491.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals?. *Evolution*, **35**, 124–138.

Fischer MC, Foll M, Excoffier L, Heckel G (2011) Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (*Microtus arvalis*). *Molecular Ecology*, **20**, 1450–1462.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Garant D, Forde SE, Hendry AP (2007) The multifarious effects of dispersal and gene flow on contemporary adaptation. *Functional Ecology*, **21**, 434–443.

Gavrilets S, Losos JB (2009) Adaptive radiation: contrasting theory with data. *Science*, **323**, 732–737.

Grant PR, Grant BR (1992) Hybridization of bird species. *Science*, **256**, 193–197.

Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94–98.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adapta-

tion in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Johnson TC, Scholz CA, Talbot MR *et al.* (1996) Late Pleistocene desiccation of Lake Victoria and rapid evolution of cichlid fishes. *Science*, **273**, 1091–1093.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Keller I, Taverna A, Seehausen O (2011) Evidence of neutral and adaptive genetic divergence between European trout populations sampled along altitudinal gradients. *Molecular Ecology*, **20**, 1888–1904.

Konijnendijk N, Joyce DA, Mrosso HDJ, Egas M, Seehausen O (2011) Community genetics reveal elevated levels of sympatric gene flow among morphologically similar but not among morphologically dissimilar species of Lake Victoria cichlid fish. *International Journal of Evolutionary Biology*, **ID 616320**, 1–12.

Langmead B, Trapnell C, Pop M, Salzberg S (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**, R25.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Magalhaes IS, Seehausen O (2010) Genetics of male nuptial colour divergence between sympatric sister species of a Lake Victoria cichlid fish. *Journal of Evolutionary Biology*, **23**, 914–924.

Magalhaes IS, Mwaiko S, Schneider MV, Seehausen O (2009) Divergent selection and phenotypic plasticity during incipient speciation in Lake Victoria cichlid fish. *Journal of Evolutionary Biology*, **22**, 260–274.

Magalhaes IS, Lundsgaard-Hansen B, Mwaiko S, Seehausen O (2012) Evolutionary divergence in replicate pairs of ecotypes of Lake Victoria cichlid fish. *Evolutionary Ecology Research* (in press).

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Nagl S, Tichy H, Mayer WE, Takahata N, Klein J (1998) Persistence of neutral polymorphisms in Lake Victoria cichlid fish. *Proceedings of the National Academy of Sciences*, **95**, 14238–14243.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Rosenberg NA (2004) Distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Samonte IE, Satta Y, Sato A, Tichy H, Takahata N, Klein J (2007) Gene flow between species of Lake Victoria haplochromine fishes. *Molecular Biology and Evolution*, **24**, 2069–2080.

Sanetra M, Henning F, Fukamachi S, Meyer A (2009) A microsatellite-based genetic linkage map of the cichlid fish, *Astatotilapia burtoni* (Teleostei): a comparison of genomic architectures among rapidly speciating cichlids. *Genetics*, **182**, 387–397.

Schluter D (2000) *The Ecology of Adaptive Radiation*. Oxford University Press, Oxford.

Seehausen O (1996) *Lake Victoria Rock Cichlids: Taxonomy, Ecology, and Distribution*. Verduijn Cichlids, Zevenhuizen.

Seehausen O (2004) Hybridization and adaptive radiation. *Trends in Ecology and Evolution*, **19**, 198–207.

Seehausen O (2006) African cichlid fish: a model system in adaptive radiation research. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **273**, 1987–1998.

Seehausen O (2009) Progressive levels of trait divergence along a 'speciation transect' in the Lake Victoria cichlid fish *Pundamilia*. In: *Speciation and Patterns of Diversity* (eds Butlin R, Bridle J, Schluter D), p. 346. Cambridge University Press, Cambridge.

Seehausen O, Bouton N (1997) Microdistribution and fluctuations in niche overlap in a rocky shore cichlid community in Lake Victoria. *Ecology of Freshwater Fish*, **6**, 161–173.

Seehausen O, Bouton N (1998) The community of rock dwelling cichlids in Lake Victoria. *Bonner Zoologische Beiträge*, **47**, 301–311.

Seehausen O, Schluter D (2004) Male-male competition and nuptial colour displacement as a diversifying force in Lake Victoria cichlid fishes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **271**, 1345–1353.

Seehausen O, Van Alphen JJM (1999) Can sympatric speciation by disruptive sexual selection explain rapid evolution of cichlid diversity in Lake Victoria? *Ecology Letters*, **2**, 262–271.

Seehausen O, Lippitsch E, Bouton N, Zwennes H (1998) Mbipi, the rock-dwelling cichlids of Lake Victoria: description of three new genera and fifteen new species (Teleostei). *Ichtyological Exploration of Freshwaters*, **9**, 129–228.

Seehausen O, Van Alphen JJM, Witte F (1999) Can ancient colour polymorphisms explain why some cichlid lineages speciate rapidly under disruptive sexual selection? *Belgian Journal of Zoology*, **129**, 43–60.

Seehausen O, Terai Y, Magalhaes IS *et al.* (2008) Speciation through sensory drive in cichlid fish. *Nature*, **455**, 620–626.

Shriver MD, Kennedy GC, Parra EJ *et al.* (2004) The genomic distribution of population substructure in four populations using 8525 autosomal SNPs. *Human Genomics*, **1**, 274–286.

Simpson GG (1953) *The Major Features of Evolution*. Columbia University Press, New York.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology*, **57**, 758–771.

Stapely J, Reger J, Feulner PGD *et al.* (2010) Adaptation genomics: the next generation. *Trends in Ecology & Evolution*, **25**, 705–712.

Terai Y, Seehausen O, Sasaki T *et al.* (2006) Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. *PLoS Biology*, **4**, e433.

Verzijden M, Korthof R, ten Cate C (2008) Females learn from mothers and males learn from others. The effect of mother and siblings on the development of female mate preferences and male aggression biases in Lake Victoria cichlids, genus *Mbipia*. *Behavioral Ecology and Sociobiology*, **62**, 1359–1368.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 451–460.

Wagner CE, Keller I, Wittwer S *et al.* (2012) Genome-wide RAD sequence data provides unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* in press.

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology*, **14**, 611–619.

Williams L, Oleksiak M (2008) Signatures of selection in natural populations adapted to chronic pollution. *BMC Evolutionary Biology*, **8**, 282.

Wu CI (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

The authors of the paper are a team from O.S.'s research group working to apply RAD-sequence based approaches to answer questions about the history and causes of adaptive radiation in cichlid fishes. I.K. is a molecular population geneticist with particular interests in adaptation, speciation and conservation genetics. C.E.W. is an evolutionary biologist with interests in speciation and the origins of diversity, and the relationships between diversity-generating processes and macroevolutionary patterns. L.G. is a masters student working on population genetics of Lake Victoria's cichlid fishes. S.M. is a molecular laboratory technician interested in applying molecular techniques to address questions about the population genetic and phylogenetic history of East African cichlid fishes. O.M.S. is a PhD student working on the evolutionary ecology of reproductive isolation in Lake Victoria cichlid fishes. A.S. is a molecular population geneticist interested in gene flow, divergence, adaptation and speciation. S.W. is a masters student working on phylogenetic inference from next-generation sequence data. O. S. is interested in processes and mechanisms implicated in the origins, maintenance and loss of species diversity and adaptive diversity.

Data accessibility

For each *de novo* assembly, a vcf file containing the genotype calls for all SNP sites is archived in DRYAD under doi:10.5061/dryad.sr14q.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Proportion of outliers (FDR = 20%; prior odds 10) among all polymorphic SNPs between all species pairs for each of the three assemblies (see Table S1 for details on assemblies).

Fig. S2 Results of outlier scans assuming even prior odds for all pairwise comparisons among the five cichlid species.

Fig. S3 (a) Results of a STRUCTURE analysis of the full dataset of 10 663 SNPs. (b) Maximum likelihood tree based on the full dataset.

Fig. S4 Results of STRUCTURE analyses of three data subsets.

Table S1 Number of RAD loci and polymorphic sites obtained with different assembly and mapping criteria in 50 individuals from five cichlid species.

Table S2 Genetic diversity of five haplochromine species (*Mbipia lutea*, *Mbipia mbipia*, *Pundamilia nyererei*, *Pundamilia* sp. 'pink anal fin' and *Pundamilia pundamilia*) at Makobe Island, Southern Lake Victoria based on the M2 assembly.

Table S3 Neutral F_{ST} between all species pairs estimated based on 5331 intermediate SNPs (i.e. between lower and upper quartiles of a list of all SNPs arranged in order of increasing global F_{ST}).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.