MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 787–798 doi: 10.1111/mec.12023

Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation

CATHERINE E. WAGNER,*† IRENE KELLER,*† SAMUEL WITTWER,*† OLIVER M. SELZ,*† SALOME MWAIKO,*† LUCIE GREUTER,*† ARJUN SIVASUNDAR*†‡ and OLE SEEHAUSEN*†

*Department of Fish Ecology & Evolution, EAWAG Centre for Ecology, Evolution and Biogeochemistry, Seestrasse 79, 6047, Kastanienbaum, Switzerland, †Institute of Ecology and Evolution, Aquatic Ecology, University of Bern, Baltzerstrasse 6, 3012, Bern, Switzerland

Abstract

Although population genomic studies using next generation sequencing (NGS) data are becoming increasingly common, studies focusing on phylogenetic inference using these data are in their infancy. Here, we use NGS data generated from reduced representation genomic libraries of restriction-site-associated DNA (RAD) markers to infer phylogenetic relationships among 16 species of cichlid fishes from a single rocky island community within Lake Victoria's cichlid adaptive radiation. Previous attempts at sequence-based phylogenetic analyses in Victoria cichlids have shown extensive sharing of genetic variation among species and no resolution of species or higher-level relationships. These patterns have generally been attributed to the very recent origin (<15 000 years) of the radiation, and ongoing hybridization between species. We show that as we increase the amount of sequence data used in phylogenetic analyses, we produce phylogenetic trees with unprecedented resolution for this group. In trees derived from our largest data supermatrices (3 to >5.8 million base pairs in width), species are reciprocally monophyletic with high bootstrap support, and the majority of internal branches on the tree have high support. Given the difficulty of the phylogenetic problem that the Lake Victoria cichlid adaptive radiation represents, these results are striking. The strict interpretation of the topologies we present here warrants caution because many questions remain about phylogenetic inference with very large genomic data set and because we can with the current analysis not distinguish between effects of shared ancestry and post-speciation gene flow. However, these results provide the first conclusive evidence for the monophyly of species in the Lake Victoria cichlid radiation and demonstrate the power that NGS data sets hold to resolve even the most difficult of phylogenetic challenges.

Keywords: adaptive radiation, cichlid, Lake Victoria, next generation sequencing, phylogenetics, RAD-seq

Received 27 March 2012; revision received 26 June 2012; accepted 4 July 2012

Introduction

High-throughput sequencing technologies have revolutionized the questions that can be addressed and the taxa that can be studied using genome-wide approaches. The field of population genomics is rapidly expanding, and studies are now possible on

Correspondence: Ole Seehausen, Fax: +41 (0)31 631 30 08; E-mail: ole.seehausen@eawag.ch

‡Present address: National Centre for Biological Sciences, Tata Institute for Fundamental Research, GKVK Bellary Road, Bangalore, 560065, India

unprecedented scales even in non-model organisms. Rarely, however, have next generation sequence (NGS) data been applied to questions in phylogenetics (but see Emerson *et al.* 2010; Rubin *et al.* 2012). The ease and decrease in cost of generating NGS-based data sets for phylogenetic inference will make them increasingly common in the coming years, and the forthcoming availability of these data represents an exciting prospect for addressing many difficult phylogenetic problems.

Most currently published studies analysing phylogenomic-scale sequence data have focused on resolving the relationships among very divergent taxa, such as the relationships among metazoan groups (Philippe *et al.* 2009), mammalian orders (Prasad *et al.* 2008) or divergent lineages of fungi (Rokas *et al.* 2003). This bias towards the analysis of distantly related taxa is due to the restricted availability of whole genome sequences, which were previously required for assembling phylogenomic-scale data sets. The increasing accessibility of NGS-based sequencing approaches will soon make data available for many more taxa, including phylogenomic studies of recently diverged taxa with increasing levels of taxon sampling possible within these groups. However, the behaviour of such data sets in phylogenomic-scale analyses has not yet been systematically evaluated. Although the unique challenges of phylogenetic inference at genome-wide scales have begun to receive attention (Delsuc *et al.* 2005; Rannala & Yang 2008; Kumar *et al.* 2012), the small number of taxa for which phylogenomic-scale analyses have thus far been possible limits the extent to which general patterns in the behaviour of phylogenomic data sets have been assessed.

We here construct reduced representation genomic libraries using restriction-site-associated DNA markers (RAD-tags; Baird *et al.* 2008) coupled with Illumina high-throughput sequencing to study phylogenomic relationships within the fastest known vertebrate species radiation, the cichlid fish adaptive radiation in Lake Victoria, East Africa. Testing species hypotheses with genetic data and reconstructing the phylogenetic relationships between Lake Victoria cichlid species is a notoriously difficult problem due to this group's extraordinary diversity (Greenwood 1974, 1980; Seehausen 1996), extremely recent origin (<15 000 years; Johnson *et al.* 1996; Stager & Johnson 2008) and ongoing hybridization between species (Seehausen *et al.* 1997). The Lake Victoria cichlid flock consists of approximately 500 species (Genner *et al.* 2004), and genetic data suggest the flock originates either from one (Meyer *et al.* 1990; Nagl *et al.* 1998) or from hybridization between several colonizing lineages (Seehausen *et al.* 2003). Shared polymorphism due to incomplete lineage sorting (aka 'deep coalescence') is expected to be high for these closely related species (Nagl *et al.* 1998), and some authors have even used the presence of extensive shared genetic variation among morphologically defined species to question the species status of members of this radiation (Samonte *et al.* 2007). Recent work using genome-wide AFLP markers, however, shows that significant genetic variation is explained at the species level (where species are defined morphologically), refuting the idea of a panmictic flock (Bezault *et al.* 2011). Still, no prior sequence-based phylogenetic work has provided any resolution at the species level within the Lake Victoria species flock. Although microsatellite and AFLP-based studies reject the panmixia hypothesis, they do not provide resolution at the species level either.

Baird *et al.* (2008) developed a restriction-site-associated DNA (RAD-tag) sequencing approach to simultaneously detect and genotype thousands of genome-wide SNPs. This approach focuses the sequencing effort on genomic regions flanking restriction sites, thereby reducing the representation of the genome to be sequenced and increasing the number of reads obtained per locus. The RAD-tag sequencing approach has been successfully used to generate genome-wide SNP data to address questions in population genomics (Hohenlohe *et al.* 2010, 2011, 2012). In addition, two studies have thus far utilized RAD-tags in the analysis of phylogenetic questions. First, SNPs derived from a RAD-sequencing approach were used for the phylogeographic study of Emerson *et al.* (2010), where they produced a high-resolution tree for the previously difficult to resolve phylogeographic problem in question (pitcher plant mosquitoes in eastern North America). Additionally, Rubin *et al.* (2012) use simulated RAD sequence data to study the accuracy of these data in phylogenetic reconstruction for several taxa with varying population sizes and evolutionary histories (specifically, *Drosophila*, mammals and yeast).

Here, we use full sequence data from RAD loci to investigate the power of these data in testing species hypotheses and reconstructing phylogenetic relationships in the Lake Victoria cichlid adaptive radiation. We investigate these questions using 16 species from a single, well-studied island community of rocky reef cichlid fishes in Lake Victoria (Makobe Island). Because whole genome assemblies of cichlids are not yet publicly available, we call genotypes from *de novo* assemblies that do not rely on a reference genome. We construct supermatrices of sequence data by concatenating thousands of 84 base pair RAD sequences located across the cichlid genome. We then investigate the trade-offs between the number of loci included in the data set and the amount of missing data in supermatrices by evaluating the resolution of trees produced from analyses of these supermatrices. We produce phylogenetic trees of unprecedented

resolution for Lake Victoria cichlids in analyses of our largest supermatrices. As one of the first uses of full RAD sequences in empirical phylogenetic analysis, these results foreshadow the power that NGS data will have in helping to resolve even the most difficult of phylogenetic problems.

Methods

Sampling

We sampled 16 species of cichlid fishes from Makobe Island, a rocky reef in southeastern Lake Victoria, Tanzania, in 2010. The Makobe Island community is the best-studied cichlid community in Lake Victoria and has been described in terms of alpha taxonomy (Seehausen *et al.* 1998a), species abundances and microhabitat associations (Seehausen & Bouton 1997), feeding ecology (Bouton *et al.* 1997; Seehausen & Bouton 1997) and reproductive ecology (Seehausen *et al.* 1998b). Fourteen of the 16 study species are among the 15 numerically most abundant species at the site; the remaining two species we included here because they represented major additional eco-morphological types typical of the Lake Victoria radiation. Our data on relative abundances of the various species in this community are based on 11 169 data points (individual fish identified to species with precise depth record) collected by one of us (O. Seehausen) between 1993 and 2001. Our 16 species include all trophic groups and most genera commonly found in rocky shore cichlid assemblages in Lake Victoria. These include the morphologically specialized epilithic algae scraping *Neochromis rufocaudalis*, *N. gigas*, *N. omnicaeruleus* and *Paralabidochromis* sp. 'short snout scraper'; morphologically more generalized algae scraping *Mbipia mbipi* and *M. lutea*; morphologically generalized omnivorous *Paralabidochromis* sp. 'rockkribensis'; morphologically specialized insect larvae sucking *Paralabidochromis chilotes* and *Pundamilia pundamilia*; zooplanktivorous *Pundamilia nyererei* and *P.* sp. 'pink anal'; the snail crusher *Labrochromis* sp. 'stone' with heavy pharyngeal bones and enlarged molariform teeth; the large piscivorous *Harpagochromis cf. serranus* with a highly specialized predator dentition; the egg and fry eating (paedophagous) *Lipochromis melanopterus*, and '*Haplochromis*' *cyaneus* a species that uses forceps-like dentition to remove chironomid midge larvae from between filamentous algae. Finally, our collection also included an 'ecotype' of *Neochromis* that is phenotypically closely related to *N. omnicaeruleus*, from which it is differentiated in body shape and dentition but whose species/taxonomic status was less certain than that of the other species (Seehausen 1996; Magalhaes *et al.* 2012).

We chose nine to 12 individuals of each species for sequencing (Table S1, Supporting information), for a total of 156 individuals in the complete data set. All fish were identified to species based on phenotype alone 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 protocol outlined in Etter *et al.* (2011) with some modifications as described below, using an SbfI high-fidelity restriction enzyme (New England Biolabs). Each library contained between 52 and 60 individually barcoded fish, with all six base pair barcodes differing 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 (Bandelin electronic, GmbH & Co. KG) sonicator 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. All libraries were sequenced on an Illumina HiSeq 2000 platform at Fasteris (Geneva, Switzerland). Three sequencer lanes were used for the initial sequencing of the 156 individuals and another two lanes to increase the coverage of selected samples.

Quality filtering and SNP calling

Reads without the complete SbfI recognition sequence were discarded from further analyses. 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 215 million reads, each 84 base pairs in length.

All reads were pooled and used for a *de novo* assembly in ustacks (STACKS pipeline, 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 size of 125 reads (-m) and excluded all stacks with coverage lower than this threshold; lower coverage stacks may result from sequencing error and will generally provide low confidence in genotype calls. We set the maximum distance between stacks (-M) within a locus

as 2, meaning that stacks that are merged to form a locus are allowed a maximum of two base pair differences with any other stack included in the locus. Note that because this parameter constrains the number of pairwise differences between stacks, the total number of base pair differences at a locus can be higher than two when more than two stacks are merged. We repeated analyses using a second threshold for the maximum distance between stacks within a locus (-M 4), and these results produced qualitatively identical results in downstream analyses. The deleveraging algorithm of ustacks, which attempts to split loci merged incorrectly, was disabled because the maximum number of haplotypes at a given locus can be higher than a total of two (the maximum expected by the deleveraging algorithm) in our case, because we pooled reads from many individuals. We excluded putative loci with unusually high coverage (i.e. 'lumberjack stacks' of ustacks) because these loci probably derive from multiple copies of similar sequences present in the genome or from highly repetitive regions. They are thus likely to contain non-orthologous sequences. Under these parameters, the *de novo* assembly produced 89 927 loci, which corresponds well to the number of RAD loci expected based on the genome size of these species. Each SbfI restriction site may result in two RAD loci, as flanking sequence both 5′ and 3′ of the site will be sequenced.

Using the *de novo* assembly constructed in ustacks, we mapped the quality filtered, demultiplexed reads from each individual separately to the consensus sequences from these loci in bowtie v. 0.12.7 (Langmead *et al.* 2009), allowing no more than two mismatches. All reads with more than one valid alignment under these criteria were excluded.

Genotypes were called for the complete sample of individuals with Unified Genotyper from the Genome Analysis Tool kit v.1.4–19, using the SNP genotype likelihood model (GATK; DePristo *et al.* 2011; McKenna *et al.* 2010). We set parameters in GATK to only consider bases with a phred quality score of at least 20, and we used genotypes from all sites with GATK confidence scores of 10 or more. The authors of GATK only recommend use of biallelic SNP calling in this version of the program; we therefore used the program with these settings, under which a maximum of two bases are possible per site.

Data supermatrix preparation

We used python scripts to parse the files output by GATK, retaining both SNP sites and nonvariable sites, thus providing full sequence data for each sequenced fragment. Because models of molecular evolution used in maximum likelihood-based phylogenetic inference are intended for sequence data, not SNPs alone, standard phylogenetic inference methods for DNA sequence data are most appropriately used if full sequence information is gathered (e.g. both variable and nonvariable sites). We coded heterozygous sites with standard ambiguity codes. The quality-controlled, genotyped data set consists of many thousands of ~84 base pair loci per individual. Note that sites that do not pass the genotyping quality threshold are excluded, thereby sometimes reducing the length of a locus beyond the raw quality-trimmed, demultiplexed length of 84 base pairs. Due to the stochasticity inherent in amplifying and sequencing a pool of DNA fragments, the number of reads varied across individuals and loci, and thus there is not sequence data for every individual at every RAD site in the genome. For phylogenetic analyses, we concatenated sequence data and constructed 'supermatrices' (de Queiroz & Gatesy 2007) by inserting missing data symbols into the data matrix for loci without data for a given individual.

We were interested in assessing how increasing the number of loci included in the data set, which also entails increasing the amount of missing data in the supermatrix, influences the outcome of phylogenetic inference. We prepared seven supermatrices to assess this trade-off. At one extreme, we used a threshold of 145 of the 156 individuals in the data set with sequence data at a given locus (e.g. allowing a maximum of 11 individuals to have missing sequence data at any given locus); all loci with sequence data for fewer than 145 individuals were excluded. We refer to the threshold for the minimum number of individuals with sequence data required to include a locus in the supermatrix as 'min individuals' from here onward. The other 'min individuals' values we assessed were 125, 115, 110, 100, 75 and 15 individuals per locus. Lowering this threshold increases the number of base pairs included in the analysis but, at the same time, increases the amount of missing data in the supermatrix.

Phylogenetic methods

We inferred phylogenies using each of the seven data supermatrices discussed earlier. Because of its ability to efficiently handle very large data sets, we used a maximum likelihood approach in RAxML 7.2.8 for phylogenetic analyses (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, for single full ML tree searches, and 100 replicates of RAxML's rapid bootstrap algorithm to account for uncertainty in the estimation of the topology (Stamatakis *et al.* 2008).

Results

We obtained a total of 512 million 100 base pair reads from five Illumina lanes. Because of differences in the number of reads obtained among Illumina runs, and the fact that some individuals were repeated in multiple lanes, the total number of reads per individual in the data set varied from 8947 to more than 6.5 million after quality filtering (Fig. S1, Supporting information).

The *de novo* assembly of all reads produced a total of 89 927 unique contigs, or putative loci (hereafter referred to simply as loci; we assume that sequences mapping to a given contig are orthologous). Many of these loci have sequence data in only one or a few individuals: there are ~78 000 loci with sequence data for four or more individuals in the data set, and only 23 loci for which all 156 sequenced individuals have sequence data (Fig. 1a). The maximum number of loci recovered from a single individual in the data set is 59 111 and the minimum is 1844 (mean 40 349 ± 12 017; Fig. S2, Supporting information); individuals with low numbers of loci with sequence data are those individuals with relatively few total raw reads (Fig. S1, Supporting information). The majority of loci in each supermatrix retained their full 84 base pair length after genotype quality filtering in GATK (minimum 73% of loci in min individuals 145; maximum 87% of loci in min individuals 15 and 75). No locus in any of the analysed data sets was shorter than 65 base pairs in length.

The size of the supermatrix does not decrease linearly as the minimum number of individuals with sequence data per locus (i.e. 'min individuals') increases (Fig. 1a); this is because of the uneven distribution of read numbers across individuals in the data set (see Fig. S1, Supporting information) and because the finite number of RAD sites in the genome places a cap on the number of loci that can be sequenced in high-coverage individuals. The seven supermatrices we prepared for phylogenetic analysis ranged from 33 953 base pairs in width (min individuals 145) to more than 5.8 million base pairs in width (min individuals 15) (Table 1) and the proportion of missing data in the matrix ranges from 0.036 (min individuals 145) to 0.426 (min individuals 15) (Fig. 1b). The number of polymorphic sites increases with the size of the supermatrix, ranging from 5660 (min individuals 145) to 180 450 (min individuals 15) (Fig. 1c).

Phylogenetic analysis using the smallest data supermatrix (min individuals 145; 33 953 base pairs in width; Table 1a) produced topologies with very low bootstrap support, and in this analysis individuals of the same species do not form clades (Fig. 2a). This pattern changes dramatically as more data are added to the analysis. The phylogenetic analysis with the supermatrix of 370 024 base pairs in width (min individuals 125; Table 1b) shows clear species-specific clades with few cases of intermixing of individuals of different phenotypically assigned species; the majority of these clades have bootstrap support above 60% (Fig. 2b). As the data matrix increases in size, resolution of these species clades increases, as does bootstrap support on the internal branches of the tree (Figs 2 and 4). In supermatrices 3 million base pairs and greater in width, all species are reciprocally monophyletic with bootstrap support above 96%, with two exceptions. One is the sister species *Neochromis omnicaeruleus* and *N.* sp. 'unicuspid scraper',

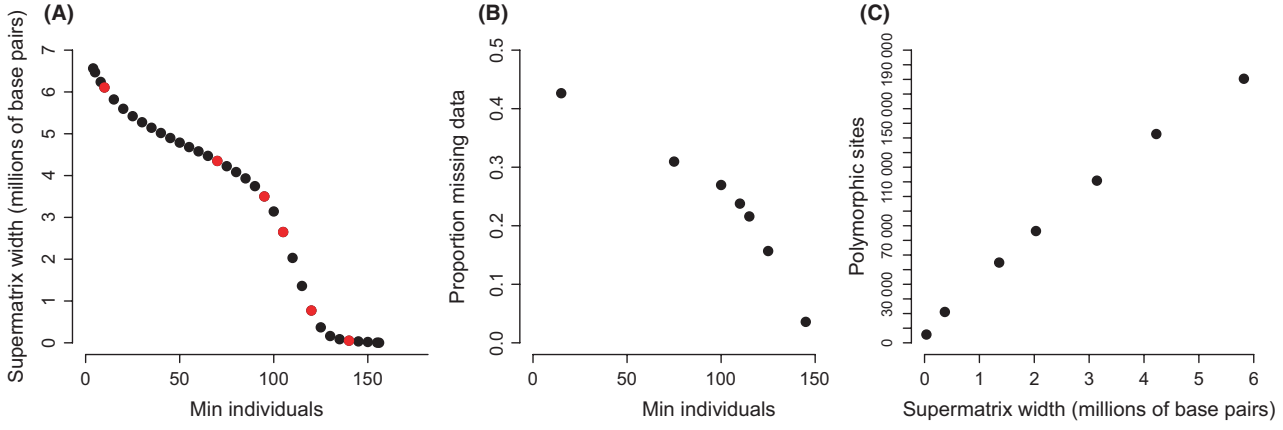

Fig. 1 Supermatrix size, missing data and genetic diversity in RAD-tag data for 156 individuals from 16 sympatric Lake Victoria cichlid species. (a) Supermatrix size decreases as we increase the minimum number of individuals with sequence data required for a locus to be included in the data set (i.e. 'min individuals'). The relationship is nonlinear due to the uneven distribution of raw sequence reads among individuals included in the study, and because the finite number of RAD sites in the genome places a cap on the number of loci that can be sequenced in high-coverage individuals (see Figs S1 and S2, Supporting information). Red dots indicate supermatrices used in phylogenetic analyses in this study (see Table 1). (b) The proportion of missing data in supermatrices decreases as the minimum individuals threshold increases. (c) The total number of polymorphic sites increases as the width of the supermatrix increases.

Table 1 The number of base pairs, number of loci and number of variable loci included in supermatrices analysed in this study. 'Min Individuals' refers to the minimum number of individuals with sequence data at a given locus required to include that locus in the corresponding supermatrix

	Min individuals	*n* base pairs	*n* loci	*n* variable loci
a	145	33 953	406	406
b	125	370 024	4418	4095
c	115	1 360 153	16 233	14 580
d	110	2 030 538	24 230	21 516
e	100	3 141 515	37 476	32 886
f	75	4 224 048	50 380	43 618
g	15	5 820 379	69 426	56 385

where each species clade has bootstrap support of 56–81% (depending on the analysis), but the clade including both of these species uniformly has 100% bootstrap support (Figs 2e,f and 3). The other exception is individuals identified as *Harpagochromis cf. serranus*, where clade support is only 66% for all individuals in the largest supermatrix (min individuals 15). However, in this case, a subclade excluding two individuals has support of 98% (Fig. 3). The relatively low bootstrap support for the monophyly of this species is not supported by other analyses; the clade including all individuals of *H. cf. serranus* has 99% or higher support in all other analyses for supermatrices >3 million base pairs in width (Fig. 2e,f).

The total number of well-supported branches on the tree generally increases as a negative exponential function of the number of base pairs in the data matrix (Fig. 4). One exception is that in analyses from the largest supermatrix (min individuals 15; Fig. 3), the number of branches with 95% or greater support slightly decreases. The curve describing the number of branches resolved with 95% bootstrap appears to asymptotically approach the value representing the total number of internal branches on the tree. The number of internal branches with high support increases markedly with increasing supermatrix size. In the analysis with data from the second-largest supermatrix (min individuals 75; Fig. 2f), all internal branches have bootstrap support of 53% or more; in the largest supermatrix (min individuals 15), all internal branches have support of 86% or more (Fig. 3).

Discussion

We find unprecedented resolution in a phylogeny of Lake Victoria cichlids when using supermatrices consisting of millions of base pairs of genome-wide sequence data. Unlike previous attempts at phylogenetic reconstruction in Lake Victoria cichlids using sequence-based methods, we find support for the reciprocal monophyly of every morphologically defined species in the data set. Individuals from all 16 species were collected in complete sympatry at a single island in Lake Victoria. In our best-supported trees, all species form monophyletic groups with 99–100% bootstrap support, with the exception of one species pair which is known to be only weakly differentiated (see discussion below); in this case, the pair forms a clade with 99–100% bootstrap support and support for each species is 59–81% (depending on the analysis considered).

The resolution of the inferred tree topology increases dramatically as the data matrix increases in size, despite the concomitant increase in missing data (Figs 2 and 4). For the smallest supermatrix (~30 000 base pairs in width; min individuals 145, Table 1a), there is essentially no resolution in the inferred phylogeny; individuals of the same species do not form clades. In contrast, in all analyses based on supermatrices >300 000 base pairs in width, individuals of the same species form generally well-supported, and reciprocally monophyletic, clades. Support for the monophyly of species increases further with increasing supermatrix width, and in the largest supermatrices, most internal branches also have high bootstrap support (Figs 2 and 3). Although the number of branches with conventionally used thresholds for reasonable bootstrap support (e.g. 60–80%) continue to increase as the supermatrix size increases, the number of branches with 95% bootstrap support is maximized in the supermatrix that is ~4.2 million base pairs in width (>1 million base pairs less than the largest supermatrix; min individuals 75) (Fig. 4). This plateau in the number of very highly resolved branches is perhaps related to the total number of species clades (16) plus internal branches connecting those clades (14) in the tree. If species represent panmictic populations, high resolution of relationships within species groups would not be expected, and as such the 29 species plus internal branches on the tree should represent an upper bound for the resolution of the tree. This pattern is thus additional evidence for genetic discontinuity across species boundaries.

One species pair, *Neochromis omnicaeruleus* and *Neochromis* sp. 'unicuspid scraper', consistently has very high bootstrap support (99–100% in analyses from the largest supermatrices). Despite perfect sorting of morphologically identified individuals into separate clades in the topologies recovered by all full ML tree searches on supermatrices >2 million base pairs in width (Fig. 2), bootstrap support for the monophyly of each of these species is consistently lower than that for other species in the analysis (59–60% for min individuals 75, compared to 99–100% for all other species; 81% for min individuals 15; Figs 2 and 3). The relatively lower

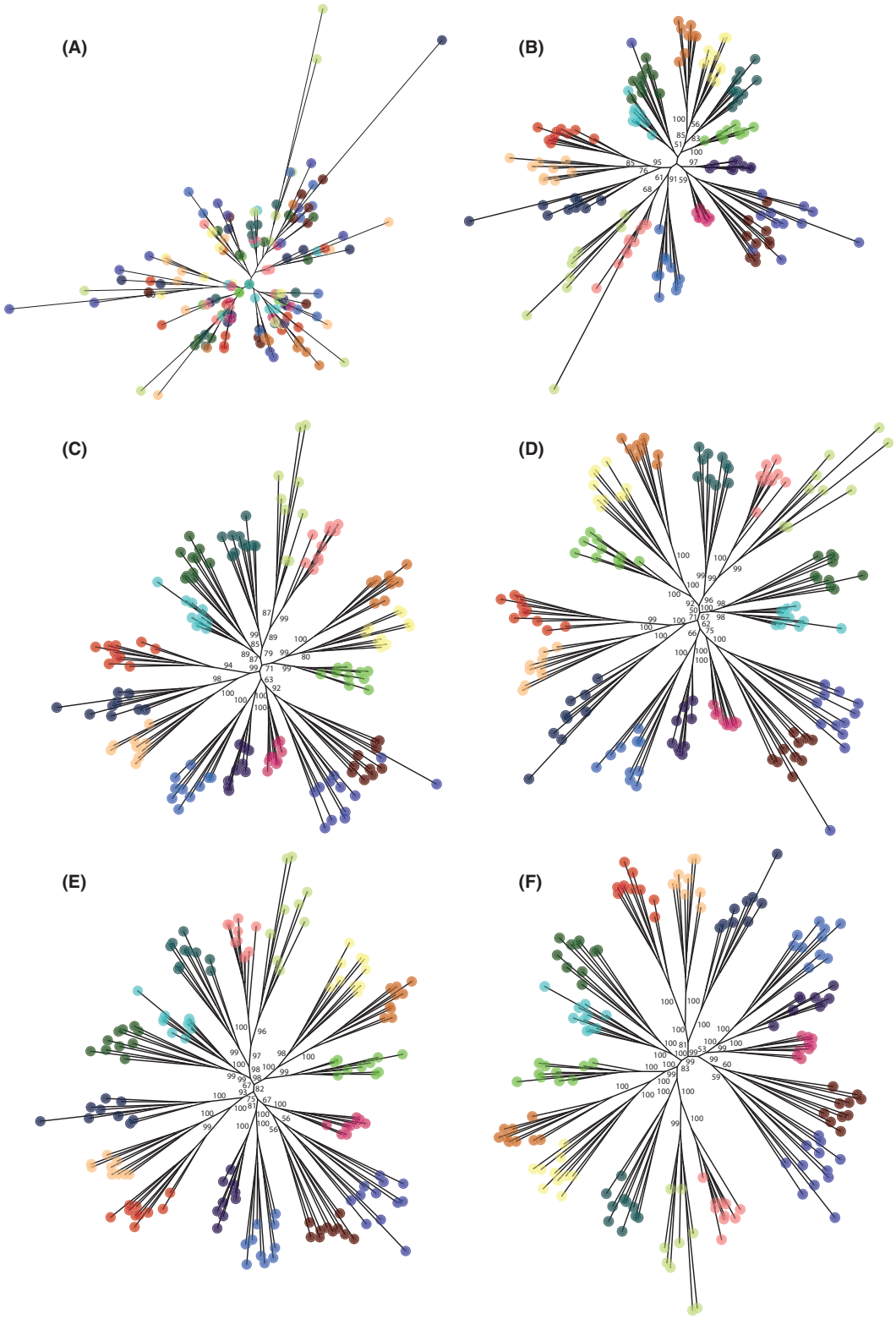

Fig. 2 Tree resolution increases as data are added to the phylogenetic analysis. Panel letters (a–f) correspond to letters identifying the supermatrices in Table 1 (see Fig. 3 for the tree corresponding to Table 1 supermatrix g). Tip colours represent sampled species (see Fig. 3 for key). Values on branches are bootstrap support from 100 rounds of bootstrapping using RAxML's rapid bootstrap algorithm. Topologies shown are the best tree from a full ML search. Bootstrap support values below 50, and all values within species groups, are not shown.

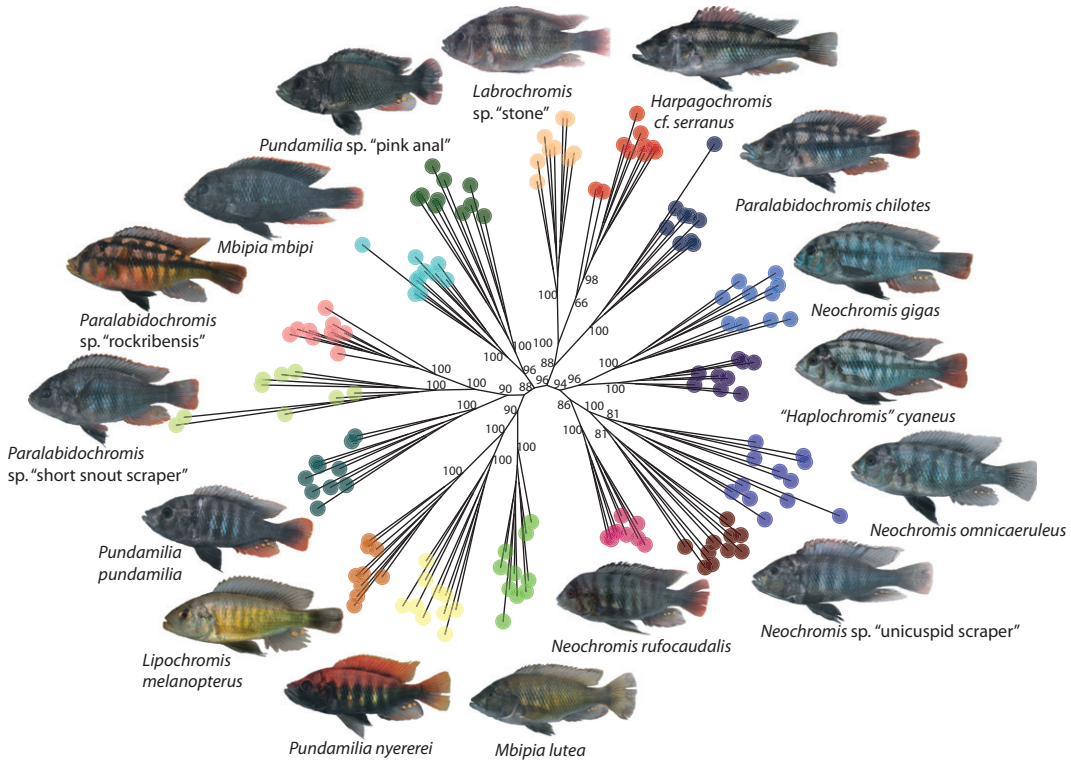

Fig. 3 The phylogeny produced based on the largest supermatrix analysed, which contains a minimum of 15 individuals out of the total 156 with sequence data per locus ('min individuals 15'; Table 1g).

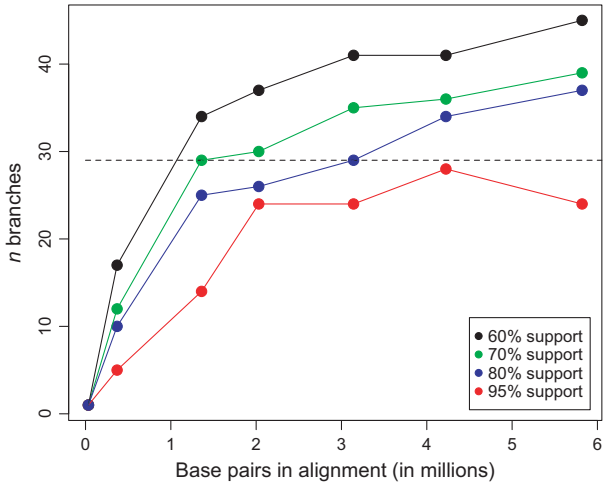

Fig. 4 The resolution of phylogenetic trees increases as a function of data matrix size. The colours indicate the proportion of branches with bootstrap support above a given threshold: black ≥ 60%; green ≥ 70%; blue ≥ 80%; red ≥ 95%. The dotted line corresponds to the total number of species + internal branches on the tree. The number of branches with 95% support plateaus close to the total number of species level and above branches on the tree.

support for the reciprocal monophyly of these species is expected given previous data showing that their divergence at neutral markers is very weak. Microsatellite data reveal a barely significant F_{ST} value of 0.01 (Magalhaes *et al.* 2012).

The high resolution of morphologically identified species in the trees produced in this study is remarkable given the very recent divergence between these species, and the inability of previous individual-based genetic studies to differentiate species in this group (Nagl *et al.* 1998; Samonte *et al.* 2007). Whereas several previous studies on selected sympatric species pairs rejected the null hypothesis of genetic panmixis using allele frequency data (AFLPs: Konijnendijk *et al.* 2011; microsatellites: Magalhaes *et al.* 2009; mitochondrial sequences: Mzighani *et al.* 2010; microsatellites: Seehausen *et al.* 2008), none of these studies were able to recover species structure in individual-based trees, clustering, or assignment tests. Microsatellite-based studies generally show small but significant F_{ST}s between the species pairs that have been studied, speaking to the very recent divergence of these species (Seehausen *et al.* 2008; Magalhaes *et al.* 2009, 2012). Although studies of larger numbers of species using AFLPs suggest that significant genetic variation within the Lake Victoria radiation is explained by phenotypically defined species (Bezault *et al.* 2011), no prior phylogenetic studies have produced evidence for species monophyly within the Lake Victoria radiation. The difficulty of the phylogenetic problem at hand

is also illustrated by our finding that the species still cannot be differentiated in analyses based on the supermatrix >30 000 base pairs in width. A recently published tree based on 654 polymorphic AFLP loci produced a similarly unresolved topology (Konijnendijk *et al.* 2011). However, the fact that tree resolution improves massively as the supermatrices grow highlights the great promise of NGS data sets for resolving difficult phylogenetic problems, perhaps particularly with regard to species delimitation. The high support for reciprocal monophyly of the species studied here conclusively supports the morphological descriptions of these species (Seehausen *et al.* 1998a).

We recover strong support for some, but not all, expected genus-level relationships in these analyses. All species of the genus *Neochromis* form a clade with high bootstrap support (99%) in trees derived from our largest supermatrices, and this clade consistently includes the species '*Haplochromis*' *cyaneus*. This species could not taxonomically be assigned to any described genus because it exhibits a mix of traits from other genera and unique traits: it resembles members of the genus *Paralabidochromis* by external appearance, but shares dentition characters with *Neochromis* and is distinct from both of these genera in scale and squamation traits (Seehausen *et al.* 1998a). The three species of the genus *Paralabidochromis* that we sampled are consistently recovered in two separate strongly supported subclades. This is not inconsistent with morphological data that suggested paraphyly of this genus (Seehausen 1996; Seehausen *et al.* 1998a). There is strong support for a clade that includes all *Pundamilia* species in the data set (83% in analyses from supermatrix min individuals 75, Fig. 2f; 90% in analyses from supermatrix min individuals 15, Fig. 3). However, this clade also includes species of the genera *Mbipia*, *Lipochromis* and *Paralabidochromis*. The following additional relationships were expected based on morphology: (a) *N. omnicaeruleus*/*N.* sp. 'unicuspid scraper'; (b) *N. omnicaeruleus* + *N.* sp. 'unicuspid scraper'/*N. rufocaudalis*; (c) *Pundamilia pundamilia*/*P. nyererei*; (d) *P. pundamilia* + *P. nyererei*/*P.* sp. 'pink anal'; (e) *Para.* sp. 'rockkribensis'/*Para.* sp. 'short snout scraper'; (f) *Mbipia mbipi*/*M. lutea* (Seehausen 1996; Seehausen *et al.* 1998a). We recover strong support for a (see discussion above) and e, and mixed support for b (86% support in the largest supermatrix; in other analyses substantially lower). We do not find support for c, d or f.

Despite the high resolution of the phylogenetic trees inferred in this study, these results should be viewed with several qualifiers and future work should critically examine their accuracy. First, our approach here of concatenating sequence data from tens of thousands of genome-wide loci does not account for the individual history of these loci and the variation therein. It is well known that gene histories differ from species histories, and concatenation approaches in some cases do not produce accurate species trees (Kubatko & Degnan 2007), and rely on fundamentally incorrect assumptions (Rannala & Yang 2008; Degnan & Rosenberg 2009). These problems are probably exacerbated where hybridization between species is present. Promising methods for explicitly and simultaneously estimating species trees given the history of gene trees exist (Edwards *et al.* 2007; Heled & Drummond 2010), but these methods cannot yet handle the computational challenge of large NGS data sets. Furthermore, although these methods in many cases effectively deal with problems arising from gene tree/species tree conflict due to incomplete lineage sorting, they can still suffer from inaccuracy when divergence is recent and population sizes are large (Leaché & Rannala 2011), and research on methods that account for horizontal gene transfer (e.g. hybridization, in the context of our study) in species tree estimation is just beginning (e.g. Kubatko 2009; Yu *et al.* 2011).

Furthermore, as data are added to phylogenetic analyses, error due to sampling variance will decrease but systematic error remains and is even compounded, as data sets become very large, creating the potential for highly supported, but incorrect, phylogenetic tree topologies (Delsuc *et al.* 2005; Jeffroy *et al.* 2006; Rannala & Yang 2008; Philippe *et al.* 2011; Kumar *et al.* 2012; Rubin *et al.* 2012). All models of molecular evolution are inherently a simplification of the actual complexity of the evolutionary process, but in some cases, failure to account for a particular feature of the evolutionary process can produce systematic bias in outcomes of phylogenetic inference (Phillips *et al.* 2004; Jeffroy *et al.* 2006). One source of model misspecification common in analyses of concatenated sequence data is the implicit assumption of rate homogeneity across the sampled genes (Rannala & Yang 2008). Analyses partitioning sequence data by gene and/or codon position have been shown to perform well (Nishihara *et al.* 2007; Rodriguez-Ezpeleta *et al.* 2007). However, it is currently unclear how to implement partitioning schemes on NGS phylogenetic data sets, because partitioning by locus would increase the parameters of ML models enormously due to the very large number of loci, and in the absence of a reference genome, loci are anonymous and thus their coding status is unknown.

Finally, concatenation of large amounts of sequence data has been shown to produce trees with very high bootstrap support values (e.g. Rokas *et al.* 2003), and the addition of sequence data is, in general, known to increase the resolution of phylogenetic trees (Alfaro

et al. 2003; Delsuc *et al.* 2005). Although nonparametric bootstrapping in the context of phylogenetics is thought to be unbiased (Efron *et al.* 1996), many authors have shown that bootstrap values underestimate confidence in tree topologies (e.g. Alfaro *et al.* 2003; Erixon *et al.* 2003). As such it is common to assume that 70% bootstrap support represents strong confidence in a clade on a tree (following from Hillis & Bull 1993). However, Taylor & Piel (2004) show that for the large, genomic-scale data set of Rokas *et al.* (2003), bootstrap values are not underestimates of clade support and suggest the possibility that for large data sets, bootstrap values lower than 95% should not be held as adequate support. Although the generality of this pattern has not been explored, this result suggests that 'rule-of-thumb' assessment of confidence in tree topologies may need to be recalculated when dealing with large data sets.

These potential sources of inaccuracy in phylogenetic inference based on genomic-scale data sets point to many possible extensions to this work as novel methods and additional data become available. Reference genomes will soon be available for this cichlid group, allowing a comparison of the results based on *de novo* assembly reported here to analyses based on alignments to the reference genome. In particular, the use of the reference genome is expected to decrease errors in the assessment of gene orthology. In the current analysis, we excluded putative loci with unusually high read numbers, which are potentially indicative of repetitive DNA or multi-copy genes, and we excluded reads mapping to more than one of the loci identified in the *de novo* assembly. Both of these steps should increase the accuracy of our assessment of sequence orthology. However, inaccurate assessment of orthology is still a likely source of error, and one which is notoriously difficult to address in phylogenomic-scale studies (Philippe *et al.* 2011). In addition, reference genome information will allow genomic mapping of loci, and with this information, partitioning analyses by linkage group, coding and noncoding portions of the genome, or other genomic subregions would relax the assumption of evolutionary model homogeneity across the genome and probably improve phylogenetic accuracy (Rannala & Yang 2008).

The data set and analyses we provide here are a novel step forward in the use of NGS data to address questions in phylogenetic reconstruction. We show that it is possible to assemble genome-wide NGS sequence data from RAD-tags into phylogenetic supermatrices containing millions of base pairs of sequence data from tens of thousands of loci, all without the use of a reference genome. We produce remarkably resolved phylogenetic trees from a species group that presents a tremendous phylogenetic challenge due to its recent origin, very large species numbers and the existence of ongoing hybridization. Our finding that these sympatric Lake Victoria cichlid species form strongly supported, reciprocally monophyletic groups highlights the power that NGS-based data sets hold for resolving species boundaries and relationships, particularly in groups with challenging evolutionary histories.

Acknowledgements

We thank Ben Rubin and Dan Rabosky for helpful discussions and comments on the manuscript, the Fish Ecology group at EAWAG for discussions and feedback, and Rémy Bruggmann and Stefan Zoller for computational assistance. Bioinformatics support was provided by the Genetic Diversity Center (GDC) at ETH Zürich and the Institute of Ecology and Evolution, University of Bern. Thanks to Rahel Thommen for assistance with laboratory work, and Mhoja Kayeba, Mohammed Haluna, Martine Maan, Erwin Ripmeester and Dora Selz-Affolter for assistance with field collecting work. 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. The work was supported through Swiss National Science Foundation grant 31003A-118293 to O. Seehausen.

References

Alfaro ME, Zoller S, Lutzoni F (2003) Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence *Molecular Biology and Evolution*, **20**, 255–266.

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

Bezault E, Mwaiko S, Seehausen O (2011) Population genomic tests of models of adaptive radiation in Lake Victoria region cichlid fish. *Evolution*, **65**, 3381–3397.

Bouton N, Seehausen O, van Alphen JJM (1997) Resource partitioning among rock-dwelling haplochromines (Pisces: Cichlidae) from Lake Victoria. *Ecology of Freshwater Fish*, **6**, 225–240.

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

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, **24**, 332–340.

Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, **6**, 361–375.

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.

Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 5936–5941.

Efron B, Halloran E, Holmes S (1996) Bootstrap confidence levels for phylogenetic trees. (vol 93, pg 7085, 1996). *Proceedings of the National Academy of Sciences of the United States of America*, **93**, 13429–13434.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 16196–16200.

Erixon P, Svennblad B, Britton T, Oxelman B (2003) Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, **52**, 665–673.

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

Genner MJ, Seehausen O, Cleary DFR *et al.* (2004) How does the taxonomic status of allopatric populations influence species richness within African cichlid fish assemblages? *Journal of Biogeography*, **31**, 93–102.

Greenwood PH (1974) Cichlid fishes of Lake Victoria, East Africa: the biology and evolution of a species flock. *Bulletin of the British Museum (Natural History) Zoology series*, **6**, 1–134.

Greenwood PH (1980) Towards a phyletic classification of the 'genus' *Haplochromis* (Pisces, Cichlidae) and related taxa. II. The species from Lakes Victoria, Nabugabo, Edward, George and Kivu. *Bulletin of the British Museum (Natural History) Zoology series*, **39**, 1–101.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, **27**, 570–580.

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, **42**, 182–192.

Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation 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.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 395–408.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? *Trends in Genetics*, **22**, 225–231.

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.

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*, **2011**, 12.

Kubatko LS (2009) Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology*, **58**, 478–488.

Kubatko LS, Degnan JH (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, **56**, 17–24.

Kumar S, Filipski AJ, Battistuzzi FU, Pond SLK, Tamura K (2012) Statistics and truth in phylogenomics. *Molecular Biology and Evolution*, **29**, 457–472.

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

Leaché AD, Rannala B (2011) The accuracy of species tree estimation under simulation: a comparison of methods. *Systematic Biology*, **60**, 126–137.

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*, **14**, 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.

Meyer A, Kocher TD, Basasibwaki P, Wilson AC (1990) Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial-DNA sequences. *Nature*, **347**, 550–553.

Mzighani SI, Nikaido M, Takeda M *et al.* (2010) Genetic variation and demographic history of the Haplochromis laparogramma group of Lake Victoria-An analysis based on SINEs and mitochondrial DNA. *Gene*, **450**, 39–47.

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 of the United States of America*, **95**, 14238–14243.

Nishihara H, Okada N, Hasegawa M (2007) Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biology*, **8**, R199.

Philippe H, Derelle R, Lopez P *et al.* (2009) Phylogenomics revives traditional views on deep animal relationships. *Current Biology*, **19**, 706–712.

Philippe H, Brinkmann H, Lavrov DV *et al.* (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology*, **9**, e1000602.

Phillips MJ, Delsuc F, Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, **21**, 1455–1458.

Prasad AB, Allard MW, Green ED, Program NCS (2008) Confirming the phylogeny of mammals by use of large comparative sequence data sets. *Molecular Biology and Evolution*, **25**, 1795–1808.

de Queiroz A, Gatesy J (2007) The supermatrix approach to systematics. *Trends in Ecology & Evolution*, **22**, 34–41.

Rannala B, Yang Z (2008) Phylogenetic inference using whole Genomes. In: *Annual Review of Genomics and Human Genetics*, pp. 217–231.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B *et al.* (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology*, **56**, 389–399.

Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

Rubin BER, Ree R, Moreau CS (2012) Inferring phylogenies from RAD sequence data. *PLoS ONE*, **7**, e33394.

Samonte IE, Satta Y, Sato A *et al.* (2007) Gene flow between species of Lake Victoria haplochromine fishes. *Molecular Biology and Evolution*, **24**, 2069–2080.

Seehausen O (1996) Lake Victoria Rock Cichlids. Verduijn Cichlids, Zevenhuizen, The Netherlands.

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, Van Alpen JJM, Witte F (1997) Cichlid fish diversity threatened by eutrophication that curbs sexual selection. *Science*, **277**, 1808–1811.

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

Seehausen O, Witte F, van Alpen JJM, Bouton N (1998b) Direct mate choice maintains diversity among sympatric cichlids in Lake Victoria. *Journal of Fish Biology*, **53**, 37–55.

Seehausen O, Koetsier E, Schneider MV *et al.* (2003) Nuclear markers reveal unexpected genetic variation and a Congolese-Nilotic origin of the Lake Victoria cichlid species flock. *Proceedings Of The Royal Society Of London Series B-Biological Sciences*, **270**, 129–137.

Seehausen O, Terai Y, Magalhaes IS *et al.* (2008) Speciation through sensory drive in cichlid fish. *Nature*, **455**, 620–U623.

Stager JC, Johnson TC (2008) The late Pleistocene desiccation of Lake Victoria and the origin of its endemic biota. *Hydrobiologia*, **596**, 5–16.

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.

Taylor DJ, Piel WH (2004) An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data. *Molecular Biology and Evolution*, **21**, 1534–1537.

Yu Y, Cuong T, Degnan JH, Nakhleh L (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Systematic Biology*, **60**, 138–149.

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 Lake Victoria's cichlid fishes. 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. I.K. is a molecular population geneticist with particular interests in adaptation, speciation and conservation genetics. S.W. is a masters student working on phylogenetic inference from next-generation sequence data. O.M.S. is a PhD student working on the potential for interspecific hybridization to generate functional novelty and the formation of new species. 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. L.G. is a masters student working on population genetics of Lake Victoria's cichlid fishes. A.S. is a molecular population geneticist interested in gene flow, divergence, adaptation and speciation. O.S. is interested in processes and mechanisms implicated in the origins, maintenance and loss of species diversity and adaptive diversity.

Data accessibility

All supermatrices analysed in this study, and the VCF file containing the genotype calls from which supermatrices were assembled, are archived in DRYAD under doi: 10.5061/dryad.6300n.

Supporting information

Additional Supporting Information may be found in the online version of this article.

Table S1 Species sampled, and voucher specimen ID numbers.

Fig. S1 Read numbers per individual after quality filtering. Colours correspond to species colours used throughout the study; each bar is one individual. Variation in read numbers derives from stochasticity in the volume of data generated from individual Illumina lanes and from repeat coverage of some individuals in multiple Illumina lanes.

Fig. S2 Total number of loci with sequence data per individual in the data set. Colours correspond to species colours used throughout the study; each bar is one individual. Individuals with very low numbers of loci are those individuals with low numbers of raw reads from Illumina sequencing (see Fig. S1, Supporting information).

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.