The Genomic Substrate for Adaptive Radiation: Copy Number Variation across 12 Tribes of African Cichlid Species

Joshua J. Faber-Hammond[1], Etienne Bezault[2], David H. Lunt[3], Domino A. Joyce[3], and Suzy C.P. Renn[1,*]

[1]Department of Biology, Reed College, Portland OR 97202

[2]BOREA Research Unit, MNHN, CNRS 7208, Sorbonne Université, IRD 207, UCN, UA, Paris, France

[3]Department of Biological and Marine Sciences, University of Hull, Hull Kingston-Upon-Hull, United Kingdom

*Corresponding author: E-mail: renns@reed.edu.

Accepted: August 27, 2019

Data deposition: Data associated with this study are available at the NCBI GEO archive under the accession GSE117914.

Abstract

The initial sequencing of five cichlid genomes revealed an accumulation of genetic variation, including extensive copy number variation in cichlid lineages particularly those that have undergone dramatic evolutionary radiation. Gene duplication has the potential to generate substantial molecular substrate for the origin of evolutionary novelty. We use array-based comparative heterologous genomic hybridization to identify copy number variation events (CNVEs) for 168 samples representing 53 cichlid species including the 5 species for which full genome sequence is available. We identify an average of 50–100 CNVEs per individual. For those species represented by multiple samples, we identify 150–200 total CNVEs suggesting a substantial amount of intraspecific variation. For these species, only ∼10% of the detected CNVEs are fixed. Hierarchical clustering of species according to CNVE data recapitulates phylogenetic relationships fairly well at both the tribe and radiation level. Although CNVEs are detected on all linkage groups, they tend to cluster in "hotspots" and are likely to contain and be flanked by transposable elements. Furthermore, we show that CNVEs impact functional categories of genes with potential roles in adaptive phenotypes that could reasonably promote divergence and speciation in the cichlid clade. These data contribute to a more complete understanding of the molecular basis for adaptive natural selection, speciation, and evolutionary radiation.

Key words: cichlid, gene duplication, genomic architecture, adaptive radiation, copy number variation.

Introduction

The most dramatic cichlid assemblages, representing the majority of the morphological, ecological, and behavioral diversity, are found among lacustrine radiations endemic to Lakes Victoria, Malawi, and Tanganyika (Fryer and Iles 1972; Turner 2007). The recent Malawi (<5 Myr isolated) and Victoria (1 Myr isolated) radiations include over 500 species each, whereas the older Tanganyikan radiation (10–20 Myr) is less speciose based on genetic (Meyer et al. 2016; Malinsky et al. 2018) and ecological data (Ivory et al. 2016) (reviewed by Salzburger [2018]). As such, this clade has appropriately been a starting point for the quest for a genomic basis of adaptive radiation. Adaptive radiation, the evolution of genetic and ecological diversity leading to species proliferation in a lineage, is classically viewed as the result of differential selection in heterogeneous environments (Dobzhansky 1937; Mayr 1963; Schluter 2000) and is studied in classic systems such as Darwin's finches (Darwin 1859), amphipods and cottoid fish in Lake Baikal (Fryer 1991), the Caribbean anoles (Losos et al. 1998), and the Hawaiian Silverswords (Baldwin and Sanderson 1998). However, both ecological factors and species-specific intrinsic traits (e.g., historical contingency, degree of plasticity, and genomic factors) influence the diversification of a lineage (Kassen 2009; Wagner et al. 2012; Hulsey et al. 2018). Advances in genomic tools now allow us to conduct the necessary investigations of genomic characteristics associated with the ability of some populations to generate such phenotypic novelty and successive speciation.

The initial sequencing of five cichlid genomes (Brawand et al. 2014) supported earlier array-based studies on a handful of species (Machado et al. 2010) that demonstrate a high rate of copy number variation among radiating cichlid lineages.

However, it is necessary to further investigate the patterns of structural polymorphism throughout the cichlid phylogeny. It is well established that gene duplication and the subsequent evolution of duplicates are an important sources of genetic (Taylor and Raes 2004) and functional novelty (Ohno 1970; Hahn 2009; Kondrashov 2012), such that 30–65% of all functional genes (Zhang 2003) are thought to have originated through fixation of gene duplicates (Han et al. 2009). Within the cichlid radiation, a number of duplicate loci are known to be involved in categories of functional/phenotypic divergence including pigmentation (Sugie et al. 2004; Braasch et al. 2006; Watanabe et al. 2007), opsins (Carleton and Kocher 2001; Spady et al. 2005; Spady et al. 2006; Terai et al. 2006; Seehausen et al. 2008), sex-determination loci (Cnaani and Kocher 2008; Shirak et al. 2008), immune function (Takahashi-Kariyazono 2017), neurohormone systems (Chen and Fernald 2006; Summers and Zhu 2008), and hox gene patterning (Santini and Bernardi 2005).

More cichlid genome-scale data are needed to build a better picture of structural evolution in this lineage. We employ interspecific array-based comparative genomic hybridization (aCGH) to interrogate structural polymorphism across the African cichlid lineage. We do so with a genome-wide high-density oligonucleotide array designed to the consensus of available cichlid genomes (Brawand et al. 2014). The technique of aCGH can identify duplicate regions that may be collapsed and therefore not detected with whole genome sequence assembly approaches (e.g., human: Locke et al. [2003] and Redon et al. [2006]; rice: Yu et al. [2013]; *Drosophila*: Dopman and Hartl [2007]; *Dictyostelium discoideum*: Bloomfield et al. [2008]; experimental evolution in yeast: Lynch et al. [2008]). By using a single species as the reference to survey a phylogenetically and ecologically broad species set, one can identify genomic regions potentially involved in an organism’s ability to inhabit a specific environment (Renn et al. 2010; Gazave et al. 2011; Gilbert et al. 2011; Skinner et al. 2014). Such an approach has been applied to identify genomic duplications and deletions within a single species, such as tissue specificity in *Chlamydia trachomatis* (Brunelle et al. 2004) and adaptation to cold in icefish (Chen et al. 2008; Coppe et al. 2013). Also, when comparing between species, aCGH can be applied to identify structural changes associated with population divergence and speciation as has been done for *Anopheles* (Turner et al. 2005; Riehle et al. 2006) and *Littorina* (Panova et al. 2014).

Here, we describe the genomic diversity with regard to DNA copy number variants (CNVs) found among the cichlid African assemblage. We first ask whether the characteristics and locations of the CNVs identified by aCGH overlap with those previously identified by depth of sequence coverage approaches, comparing assemblies based on short-read and long-read data. With these data, we are then able to demonstrate a high level of intraspecific variation relative to interspecific variation. By considering a greater number of species across the cichlid radiation, we are able to investigate the extent to which the pattern of CNVs across species reflects their phylogenetic relationships. With regard to genomic architecture, we interrogate loci for evidence of genomic hotspots and characterize variable regions with regard to transposable elements (TEs) and gene content. We identify functional categories of affected genes that are consistent with a role in adaptation to the environment. These data allow a global overview of the patterns of copy number variation among cichlids and enhance our understanding of the molecular basis for speciation and adaptive radiation.

Materials and Methods

Microarray Construction

We used a custom multispecies cichlid 135K Nimblegen array for which probes were designed based on the consensus sequence generated from the Satsuma multiple genome alignment across the five sequenced species; *Oreochromis niloticus* (Or.ni), *Astatotilapia burtoni* (As.bu), *Metriaclima zebra* (Me.ze), *Neolamprologus brichardi* (Nl.br), and *Pundamilia nyererei* (Pu.ny). The array was designed to include three probes for each annotated gene (Brawand et al. 2014) (~70.5K probes for over 24K genes) and a single intergenic probe approximately every 6 kb (~64K probes). To maximize cross species hybridization on the array, all probes were required to share at least 95% sequence identity with at least one species from each of the three sequenced lineages (i.e., Or.ni, Nl.br, and at least one of the three Haplochromine species). To minimize representation of repetitive elements, each probe was allowed a maximum of ten such matches in each species. Seventy-five percent of probes met these criteria in all five species, however, when assembly gaps interrupted the multiple sequence alignment, probes were designed based on genome sequence for the available subset of the species, and in some cases (12.3% of all probes) only Or.ni genome sequence, which is better assembled, was available as a template for probe design.

Samples

A total of 168 samples from 53 species were collected and donated from a variety of sources (supplementary table S1, Supplementary Material online) and all were previously stored in either 100% or 70% EtOH for at least 1 year. For the purposes of figures and tables, species are given a four letter code determined by genus.species (table 1). Each species is designated as belonging to a specific lake and assigned to a “radiation” based on monophyletic relationship originating from a specific biogeographic area and assigned to “tribe” as a taxonomic grouping at a level intermediate to genus and family.

Table 1

Species Code Key for Samples Used in This Study

Sp. Code[a]	*N*[b]	Genus/species	Sp. Code[a]	*N*[b]	Genus/species
Ap.al	2	*Alcolapia alcalicus*	Ne.om	3	*Neochromis omnicaeruleus*
Ap.gr	3	*Alcolapia grahami*	Nl.br	6	*Neolamprologus brichardi*
Ap.la	3	*Alcolapia latilabris*	Ny.mi	2	*Nyassachromis microcephalus*
Ap.nd	3	*Alcolapia ndalalani*	Op.ve	4	*Ophthalmotilapia ventralis*
Ao.al	3	*Astatoreochromis alluaudi*	Or.ni	6	*Oreochromis niloticus*
As.bu	8	*Astatotilapia burtoni*	Pd.to	2	*Pallidochromis tokolosh*
As.bl	3	*Astatotilapia c.f. bloyeti*	Pa.ch	3	*Paralabidochromis chilotes*
As.ca	6	*Astatotilapia calliptera*	Pa.ro	3	*Paralabidochromis sp. rockribensis*
As.fl	2	*Astatotilapia flavijosephi*	Pb.mu	3	*Pseudocrenilabrus multicolor victoriae*
Bo.mi	3	*Boulengerochromis microlepis*	Pu.ny	6	*Pundamilia nyererei*
Ca.ma	3	*Callochromis macrops*	Pg.ma	3	*Pungu maclareni*
Cc.lp	3	*Cyprichromis leptosoma*	Rh.lg	3	*Rhamphochromis longiceps*
Co.bo	2	*Copadichromis borleyi*	Sa.ga	3	*Sarotherodon galilaeus*
Co.vi	3	*Copadichromis virginalis*	Sa.kn	2	*Sarotherodon knauerae*
Cx.fu	2	*Cyathopharynx furcifer*	Sa.la	3	*Sarotherodon lamprechti*
Cp.fr	1	*Cyphotilapia frontosa*	Sa.me	3	*Sarotherodon melanotheron*
Di.gr	3	*Diplotaxodon greenwoodi*	St.ma	3	*Stomatepia mariae*
Er.cy	3	*Erectmodus cyanostictus*	Ti.de[c]	3	*Coptodon deckerti*
Hp.th	2	*Harpagochromis thereuterion*	Ti.ej[c]	4	*Coptodon ejagham*
Hm.bi	3	*Hemichromis bimaculatus*	Ti.fu[c]	2	*Coptodon fusiforme*
Hm.fa	3	*Hemichromis fasciatus*	Ti.ko[c]	1	*Cotpodon kottae*
Ju.or	3	*Julidochromis ornatus*	Ti.zi[c]	3	*Coptodon zillii*
Ko.ei	3	*Konia eisentrauti*	Tp.re	3	*Tropheops sp.* Red Cheek
La.fu	2	*Labeotropheus fuelleborni*	Tr.mo	4	*Tropheus moorii*
Lp.el	4	*Lepidiolamprologus elongatus*	Va.mo	2	*Variabilichromis moorii*
Lo.la	3	*Lobochilotes labiatus*	Xn.sp	3	Xenotilapia spiloptera
Me.ze	6	*Metriaclima zebra*			

[a] Sp. Code is an abbreviation of genus and species modified to be nonredundant between different taxonomic designations.

[b] N indicates the number of independent individuals sampled for gDNA.

[c] The genus *Coptodon* and tribe Coptodonini were previously known as *Tilapia* and Tilapiini respectively, and 4-letter species codes reflect the previous designation to prevent confusion with the genus *Copadichromis.*

Array Hybridization and Processing

For each species, DNA was extracted from muscle or fin tissue using a standard proteinase-K and phenol–chloroform protocol. After quantification and quality assessment, genomic DNA was labeled using dual-color labeling kits (NimbleGen), following manufacturer's protocol. Equal amounts of labeled products from test samples (Cy3) and reference sample (Cy5) were hybridized on the 12-plex custom Cichlid array. We used As.bu as a reference species for this study because it represents the phylogenetic barycenter of the five sequenced cichlid species that were the basis for array design, and it is intermediately positioned within the phylogeny of species we analyzed. Therefore, the use of As.bu as the reference in all hybridizations should limit the potential hybridization bias.

After 64-h hybridization at 42 °C in a NimbleGen Hybridization Station 4, arrays were washed, dried, and scanned on the GenPix 4000 Scanner at a resolution of 5 μm/pixel. Each array was scanned individually to optimize laser power and PMT. High-resolution images were processed with DEVA v.2.1 (NimbleGen). Dual-color signal-intensity matrices (GEO: GSE117914) were exported and analyzed in RStudio v3.1.3 (RStudio Team 2015) using the Ringo (Toedling et al. 2007) and limma packages (Ritchie et al. 2015).

CNV Calling

Taking advantage of the five sequenced cichlid genomes, we optimized the preprocessing and normalization pipeline by comparing copy number variation estimates from our aCGH technique to the results from previous read-depth analysis (Brawand et al. 2014) (see supplementary information, Supplementary Material online, for details). For our aCGH analysis, we exported probe-level red and green intensity data from each step in the Deva CGH pipeline prior to segmentation and tested various between-array and GC normalization algorithms in RStudio. Our optimal pipeline took qspline signal normalized data from DEVA and applied MA2C GC normalization prior to CNV calling. For segmentation, we used default parameters in DNAcopy v1.34.0 (Seshan and Olshen 2016) which implements a Circular Binary Segmentation method for calling relative copy number for genomic intervals. We filtered out all segments that were called by <3 consecutive probes based on the Or.ni assembly. To call a CNV in an individual, we required the identified segment to have a log 2 aCGH copy number ratio <-0.8 or >0.8. For each CNV, we used the BEDOPS v2.4.5 bedmap and merge tools (Neph et al. 2012) to identify and merge those that had 50% reciprocal overlap with CNVs in any other sample (supplementary table S2, Supplementary Material online). These were classified as distinct copy number variation events (CNVEs) based on the assumption that these variants are evolutionarily related.

The ±0.8 CNV-calling threshold was selected based on meta-analysis testing the strength of correlation between our aCGH data set and the read-depth-based copy number ratios for annotated genes, and examining ratios of concordant (defined as either gains or losses in both data sets) and discordant CNV calls (defined as gains in one data set and losses in the other) (fig. 1 and supplementary information, Supplementary Material online). Briefly, we tested several CNV-calling thresholds, and ±0.8 provided strong correlations with read-depth copy number ratios and filtered out the vast majority of discordant CNV calls. Increasing thresholds beyond ±0.8 greatly reduced the number of concordant candidate genes detected as copy number variable in both data sets, thereby limiting our ability to conduct inferential downstream analyses. Additionally, we used categorized concordant and discordant genes to calculate probe-level summary statistics for the different CNV subsets and performed more in-depth comparative platform analyses (supplementary information, Supplementary Material online).

To call a CNVE at the species-level, we determined median log 2 hybridization ratios across all samples of a single species. It is important to note that sample size varied across species (table 1), however this step allowed us to summarize results for all representatives of each species and balance the detection of rare alleles with identifying intraspecific species variants. Although uneven sample sizes could impact CNVE counts per species or inferred pairwise relationships, we found no correlations between sample size and CNVE counts across taxa ($R^2 = 0.0013$). Furthermore, in support of the use of log 2 hybridization ratios, mitochondrial sequence-based phylogenetic relationships (see below) were better recapitulated by the species-level CNVE data set using the median log 2 hybridization ratios than by a resampling approach (supplementary information, Supplementary Material online). Nevertheless, the majority of our analyses rely on the full CNVE data set rather than comparison between specific species thus mitigating the impact of rare allele detection or overweighting.

In order to remove CNVEs that might represent dye-bias or array artifacts, we filtered out CNVEs that showed a <-0.3 or >0.3 species-level hybridization ratio for any As.bu sample (As.bu self-hybridization aCGH data). By doing this, we focused our analysis on those CNVEs that appear copy number neutral in reference As.bu versus As.bu arrays and can therefore be accurately assessed among other species using this platform. All CNVE results reported here are relative to As.bu and not absolute copy number. Finally, we retained only those CNVEs for which the median species-level ratio surpassed the ±0.8 threshold, which were subject to downstream analyses (supplementary table S3, Supplementary Material online).

With the recent publication of the Pacific BioSciences Or.ni genome assembly (GB accession GCA_001858045.3) (Conte et al. 2017), we were able to ask whether identified copy number variable regions are more accurately assembled using long-read sequencing technology than they are in the short-read Or.ni version 2 genome assembly (GB accession GCA_000188235.2), which was used as the primary template to build our aCGH array. As liftover files are not currently available between genome builds, we performed a probe-level analysis comparing BlastN alignment results. A non-redundant set of probe sequences located within identified CNVE regions was created using BEDtools v2.24.0 intersect tool (Quinlan and Hall 2010), and command-line BlastN was run with the probe multifasta file as the query and either genome assembly as the target database using default parameters (Altschul et al. 1990; Camacho et al. 2009). Tabular output files were filtered to include perfect alignments only. For each probe, the number of perfect hits was compared between the two genome assemblies, with the prediction that true CNVE regions would map to more loci in a PacBio assembly if short-read assemblies erroneously collapse duplicates into a single locus. We repeated this analysis for 1,000 iterations using randomly selected sets of probes outside our CNVE regions to establish background levels of collapsed sequence across the genome. We performed a chi-square outlier test from the R-package "outliers" v0.14 (Dixon 1950) using the ratios of probes with more hits in the PacBio genome versus those with more hits in the Illumina genome to test whether more observed CNVE probe sequences appear resolved into multiple loci in the PacBio assembly than expected by chance. We define these ratios as "long-read assembly resolution" (LAR) indices for a given probe set.

Phylogenetic Analysis

To determine how species cluster according to CNVEs, we used RAxML v8.2.10 (Stamatakis 2014) to construct a tree based on our aCGH data. For the 1,413 species-level CNVEs,

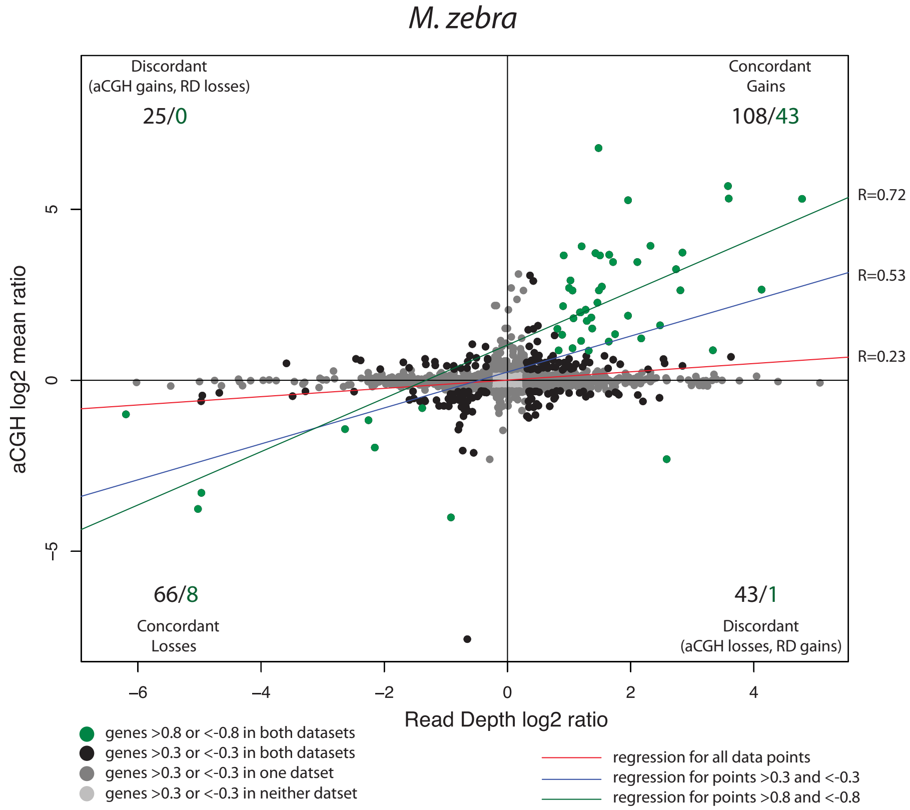

Fig. 1.—Relationship of gene log 2 ratios produced by aCGH and NGS read-depth. Each point represents the relative copy number for a single gene from the BROAD annotation. The methods share a positive correlation, confirming that they identify many of the same genomic regions as either gains or losses. The relationship is made more positive and the correlation is stronger when filtering out near-neutral CNVs, suggesting that both methods more precisely detect CNVEs with more extreme copy number ratios.

we constructed matrices for gains and losses separately coding each as “1” versus “0” for nonvariable. We then ran RAxML on the concatenated matrix using the BINGAMMA model for binary data. The best tree was determined from 1,000 algorithm iterations, and bootstrap values were assigned to each node based on an additional 100 iterations. This CNVE RAxML tree was visualized in FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and used for clustering samples in a heatmap created with the heatmap.2 tool in “gplots” v 3.0.1 (Warnes et al. 2016) in RStudio, where CNVEs were clustered using the Ward.D.2 method (Ward 1963).

To further investigate the extent to which our aCGH data recapitulate phylogenetic relationships, we created a maximum likelihood tree using RAxML for mitochondrial sequence data (*D-loop* and *ND2* sequences, Sa.kn was omitted for lack of sequence data). We aligned available *ND2* and *D-loop* using Geneious v10.2.3 (Kearse et al. 2012) for each amplicon (supplementary table S2, Supplementary Material online). We trimmed the alignments to include only overlapping sequence found in at least 90% of available sequences and concatenated the two alignments using FASconCAT v1.0 (Kück and Meusemann 2010) to run in RAxML using the GTRGAMMA model with the same number of iterations as above. This phylogenetic tree was then compared with the CNVE-based RAxML tree by calculating split distance and calculating the portion of taxa that disagree in exact placement between topologies using the Disagree tool in TOPD/FMTS v3.3 (Puigbò et al. 2007). Split distance is a metric between 0 and 1 that reflects the number of changes needed to convert one tree into the other, with values near 0 indicating the two trees are more isometric (Robinson and Foulds 1981). Additionally, we used the tanglegram tool in Dendroscope v3 visualize differences in the sequence-based and CNVE-based trees (Huson et al. 2007).

To calculate the rate of intraspecific variation captured by our arrays, we focused on variation in Me.ze, Nl.br, and Pu.ny arrays, which were among the species with the most biological replication in our experiment, having six samples from each. We created a table with the normalized log 2 copy number ratios, rather than simply gain or loss state, assigned to each CNVE for each sample (not collapsed by species). Using the same ±0.8 threshold for calling losses and gains, we categorized and sorted CNVEs according to the number of replicate samples in which it was detected (CNV calls in 1/6 to 6/6 individuals). The same heatmap.2 parameters were used to visualize whether observed intraspecific variation was an artifact of our CNV-calling threshold or reflected actual diversity in CNVEs within a species.

Genomic Architecture

To visualize the genomic distribution of CNVEs, we used the University of California-Santa Cruz genome graphs tool to produce a map on the Or.ni chromosomes for the species-level CNVEs identified in any taxa. We also mapped copy number hotspots throughout the genome using the program HD-CNV v3 (Butler et al. 2013). Due to requirements of HD-CNV, we used raw sample-level CNVs and initially collapsed only those CNVs that were exact replicates. With parameters set to assign "families" among CNVs based on 99% overlap (graphed as nodes), HD-CNV identified CNV groups with 50% reciprocal overlap (graphed as edges) that are linked to each other (analogous to CNVE definition). HD-CNV outputs were visualized using Gephi v0.9.1 (Bastian et al. 2009) as recommended. Groups consisting of ten or more CNV families were identified as "hotspots" and followed up on for downstream analyses.

To determine whether CNVEs were closely associated with transposable elements (TEs), we compared TE load between observed CNVE regions with randomly selected similar regions throughout the genome for each of six broad categories (DNA, LINE, LTR, RC, SINE, and Unknown) described in the BROAD TE data set localized in the Or.ni genome (Brawand et al. 2014) (ftp://ftp.broadinstitute.org/pub/vgb/cichlids/Annotation/TE_Annotation/). To do so, we performed binomial tests for matched pairs of genomic intervals corresponding to sets of both sample-level and species-level CNVEs. To reduce impact of the bias from array design in the selection of "random" paired intervals, we randomly selected from all possible genomic loci that share an identical number and order of genic/nongenic aCGH probes with the respective CNVE and that are approximately the same length (±10% difference). The different types of TEs were counted within sample-level and species-level CNVEs and their matched random intervals using BEDtools intersect tool, TE loads were compared within sets of matched pairs, and binomial tests were performed in Rstudio. This analysis was repeated for the 2- and 20-kb regions flanking the matched pairs to determine whether CNVEs are enriched for TEs at their boundaries, and Bonferroni corrections were applied for multiple hypothesis testing.

To address potential functional consequences of copy number variation, we first identified CNVEs as either containing or not containing any annotated genes based on the genome position using BEDtools Intersect tool. For our gene database, we used a nonredundant list of coding features compiled from Or.ni genome BROAD annotations (v2_preliminary, 2012) (Brawand et al. 2014) and supplementary homology-based predictions from GPIPE (Heger and Ponting 2007) (using Ensembl release 64 for Tetradon, Stickleback, and Human). To check whether CNVEs in cichlids were enriched for certain functional categories, we performed gene ontology (GO) enrichment on the full set of genes contained within As.bu-filtered sample-level CNVEs, species-level CNVEs, and species-level CNVEs detected within each cichlid tribe. We also tested for GO enrichment of genes in copy number hotspots detected by HD-CNV. Enrichment analyses were performed in BLAST2GO v4.1 (Gotz et al. 2008) with an false discovery rate cutoff of 0.05. For all enrichment analyses, we compared the gene test sets to the fully compiled set of annotated genes in Or.ni. A heatmap of GO term significance was constructed in Rstudio using the "gplots" R-package (Warnes et al. 2016).

Results and Discussion

Array QC

We use MA2C normalization for all analyses because it is more sensitive to detection of candidate copy number variable genes (average of 108 per species) than GCloess normalization (54 per species) and produce more uniform MA plots across all species in the study (supplementary fig. S1, Supplementary Material online). Both normalization methods give strong correlations of gene log 2 ratios, with GCloess ($R = 0.69 \pm 0.03$, $P < 0.001$) yielding a stronger correlation between aCGH and read-depth analysis than MA2C ($R = 0.55 \pm 0.06$, $P < 0.001$), however these differences are likely attributable to differences in the number of candidate genes detected by each method (supplementary information, Supplementary Material online).

Detecting Variation

Our conservative pipeline and thresholds for aCGH analysis identifies a total of 39,327 CNVs (average 234.09 ± 90.08 per sample). To better compare similar CNVs for which start and stop sites have minor variation, we define a CNVE as the region encompassed by CNVs that have 50% reciprocal overlap with each other. Merging CNVs by this criterium yields 4,428 unique CNVEs, some of which partially overlap with <50% reciprocal overlap. To account for array biases/artifacts, we also conservatively filtered out all CNVEs detected in As.bu versus As.bu arrays and with this pipeline, we retained a total of 2,879 CNVEs across all samples for downstream analysis. To present data at the species-level for all 53 species, most of which are represented by 3 or more samples, we calculate log 2 median hybridization ratio for a species for each CNVE and identify 1,413 species-level CNVEs falling beyond the −0.8 or 0.8 threshold criteria. Only one CNVE (gain in "LG8-24_24068649_24068883") was detected in a majority of species (28 of 53) suggesting that it might more

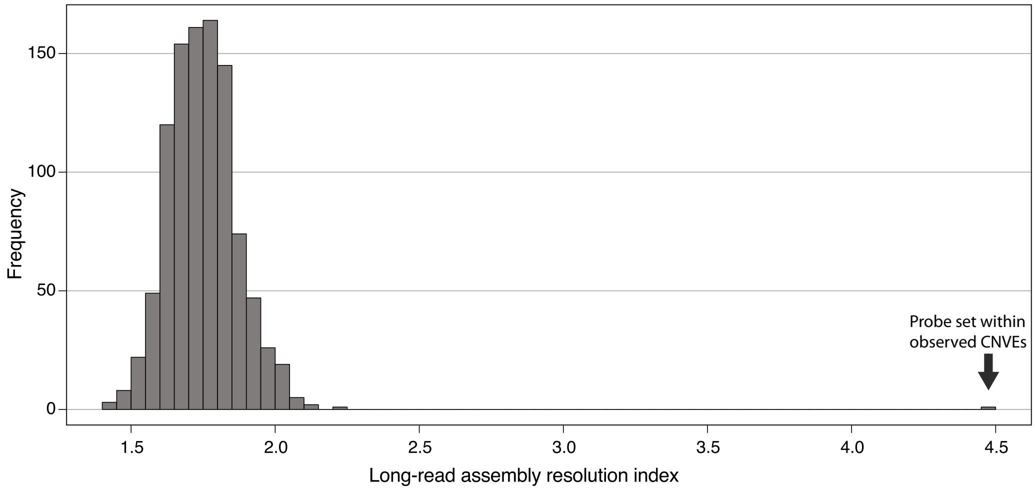

FIG. 2.—Histogram of LAR indices for probe set within observed species-level CNVEs and 1,000 iterations of randomly selected probe sets. The LAR index is defined as the ratio of counts of probes that map to more loci in the long-read PacBio assembly compared with those that map to more loci in the short-read Illumina assembly for Or.ni. Loci counts for each probe are perfect BlastN hits for probes sequences within each genome. Random probe sets were selected to have identical numbers of both exonic and noncoding probes as our observed set.

appropriately be considered to be a CNV in the reference species As.bu. All other CNVEs are identified in fewer than 50% of the species.

To determine whether our CNVEs are more accurately resolved into separate sequences in the Pacific BioSciences Or.ni genome assembly (GB accession GCA_001858045.3) (Conte et al. 2017) than in the short-read Or.ni version 2 genome assembly (GB accession GCA_000188235.2) used to build our aCGH array, CNVE probe sequences were aligned to both assemblies. In total, the 1,413 species-level CNVEs contained 13,158 nonredundant probe sequences used for BlastN alignments and 12,861 retrieved perfect alignments in both genome assemblies. From this set of probes, 1,080 (8.4%) align to more loci in the PacBio genome assembly than Illumina-based genome assembly, whereas only 242 (1.9%) align to more loci in the Illumina-based genome assembly. Therefore, our CNVE probe set has a LAR index (ratio of probes with more hits in a PacBio assembly vs. Illumina assembly) of 4.5. This ratio is greater than all 1,000 iterations performed with randomly selected probes (mean LAR = 1.75 ± 3.6E-3) and is detected as a significant outlier from the background distribution ($\chi^2 = 353.86$, $P < 2.2\text{E-}16$) (fig. 2) providing strong evidence that copy number variable regions are better assembled using long reads than short reads, and validating a portion of the CNVEs detected by our pipeline. It should be noted that we used as query the full set of CNVEs detected in any species, thus not all are expected to be of high copy number in Or.ni genomes. The instances of CNVE probes that produced more alignments within the Illumina assembly than the PacBio assembly may stem from individual variation, given that different fish were sequenced for each assembly. In addition to the Or.ni (*O. niloticus*) genome assembly, a PacBio assembly was recently published for Me.ze (*M. zebra*) (Conte and Kocher 2015). Once liftover files are made available for both, it would be of interest to do a more comprehensive analysis on the impact of sequencing technology on the ability to detect copy number variation using species-specific CNVEs identified in this study.

Quantifying Variation

We are best able to examine intraspecific variation for Me.ze, Nl.br, and Pu.ny, each of which is represented by six samples (fig. 3). Within species, we detect a total of 199 CNVEs for Me.ze (average 56.17 ± 9.2 CNVEs per sample), 255 CNVEs for Nl.br (average of 104.17 ± 17.89), and 152 CNVEs for Pu.ny (average 69.17 ± 8.28). Considering the total number of CNVEs across the genome for each species, an average of 9.06% ± 3.5 of the CNVEs are identified in all six samples suggesting that they may be fixed for that species. An average of 19.85% ± 4.79 of CNVEs are found in a majority of samples of a species, and 49.01% ± 5.13 are found in only a single sample for that species revealing a substantial amount intraspecific structural variation. The actual log 2 hybridization ratios provide additional information suggesting that some CNVE calls for an individual sample may simply fall just short of threshold criteria (fig. 3). For these CNVEs, hybridization ratios show a clear trend for directional concordance with discrete CNVE calls (gains or losses) found in other samples from the same species. However, when the threshold criteria are altered to determine whether intraspecific variation can be largely attributed to our chosen CNVE cutoff, the proportion of the six samples for a given species in which the CNVE is called remains largely unchanged (93.64% ± 1.92 at 0.6/−0.6 thresholds and 85.24% ± 1.54 at 0.3/−0.3 thresholds).

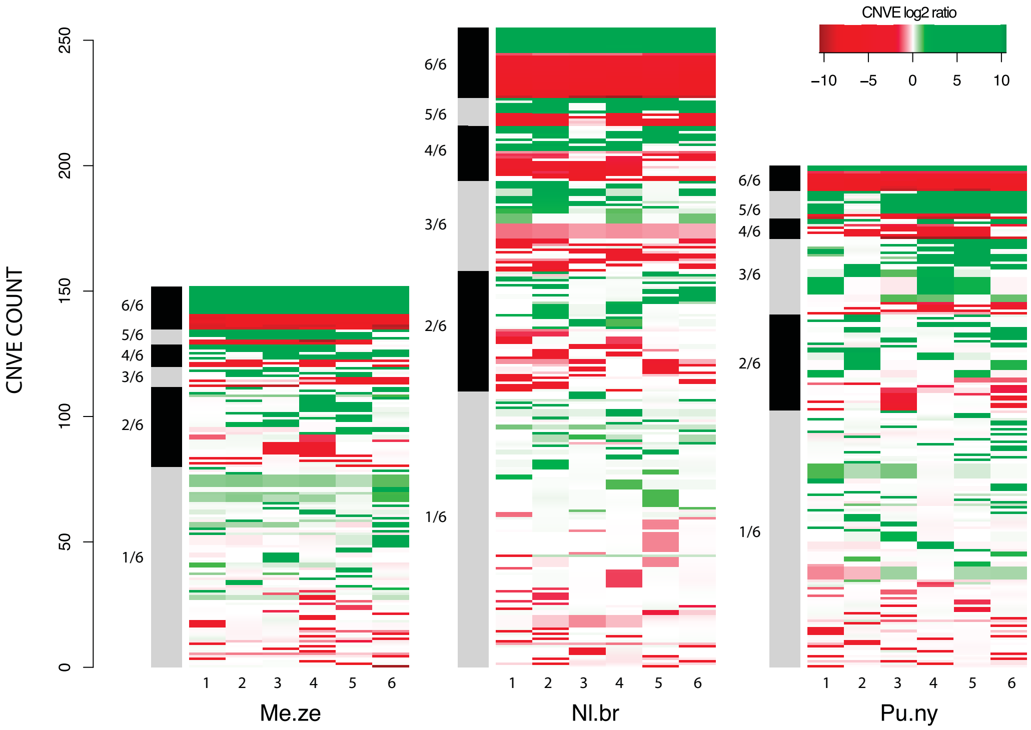

Fig. 3.—Heatmaps showing intraspecific variation in Cichlids. Six arrays were run for each of the three species shown (*Metriaclima zebra*, *Neolamprologus brichardi*, and *Pundamilia nyererei*), and heatmaps are sorted based on CNVE representation in different fractions of the six samples. The log 2 ratios highlight some clusters of CNVEs that were missed in some samples as a result of our CNV-calling thresholds of >0.8 and <−0.8, although the majority of CNVE calls appear as true individual variation based on aCGH data.

In other words, most observed individual variation does not appear to be an artifact of CNV-calling thresholds. Given this high level of individual variation noted in the species for which we had six samples, it is important to note that many of the 1,413 CNVEs called for species represented by only a few samples are likely to also represent intraspecific variation and should not necessarily be considered to be fixed for that species.

There is a wide range of CNVE sizes in the species-level data set (fig. 4). The largest detected CNVE is over 2 MB, whereas the smallest event is 67 bp. These smallest CNVEs are based on partially overlapping probes, because the array was designed in order to fit at least three probes in every gene regardless of size. Only 89 of the 1,413 species-level CNVEs are <500 bp and may not be considered CNVs by strict definitions, however we retained these copy number variable smaller regions for subsequent functional analysis due to their genic content. The mean CNVE size is ∼108 kb, whereas the median size is ∼44 kb. Detectable size for CNVEs is in part due to platform design and pipeline thresholds, making comparisons to other studies and species difficult, but for reference, among humans and nonhuman primates the median CNV size was reported to be ∼8 kb with the average human genome containing a total of 3.5+/−0.5 Mb of CNV (Sudmant et al. 2015). In total, the CNVEs from all examined species overlap with ∼58 MB of nonredundant sequence from the Or.ni genome, accounting for 6.25% of the entire assembly and suggesting a sizable portion of the genome is highly dynamic in structure across cichlids.

On average, we detect 70.94 ± 31.63 CNVEs per species (fig. 5). *Lobochilotes labiatus* (Lo.la) has the fewest CNVEs at 26, followed by *Astatotilapia calliptera* (As.ca) and *Tropheus moorii* (Tr.mo), both with 34 CNVEs, and *Paralabidochromis* sp. *rockribensis* (Pa.ro) with 35. These four species with the fewest CNVEs all belong to the tribe Haplochromini, which also contains the reference species, As.bu. Conversely, three of the four species with the most CNVEs belong to the tribe Oreochromini and show more gains than losses, which may be related to the array design biased toward Or.ni genome sequence. We find 174 CNVEs for Or.ni, 148 for *Sarotherodon knauerae* (Sa.kn), and 121 for both *Alcolapia alcalicus* (Ap.al) and *Variabilichromis moorii* (Va.mo), the latter belonging to the Lamprologini tribe. In general, at the level of tribe, the Lamprologini show the second highest number of CNVEs (average of 88.75 ± 21.82 per species) next to Oreochromini (106 ± 33.57). However, Lamprologini show an overabundance of losses compared with gains, 1.74:1, whereas, Oreochromini show an overabundance of gains compared with losses, ∼4:1. When comparing descriptive statistics at the level of tribe, it is

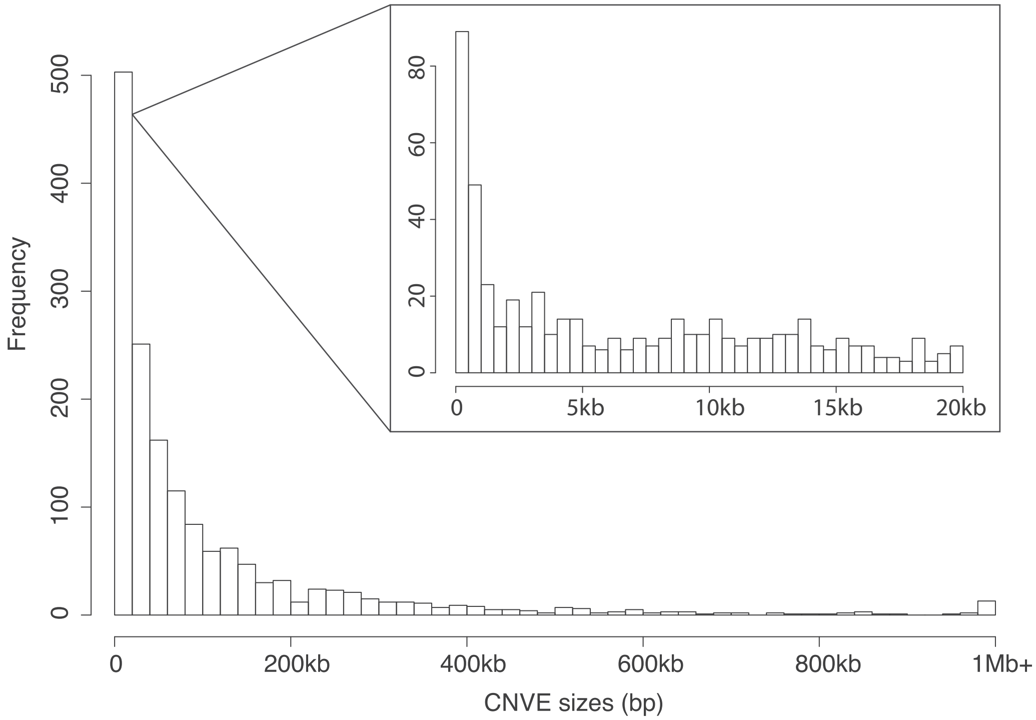

Fig. 4.—Histogram showing size distribution of CNVEs among the 53 species. Bin size in main histogram is 20 kb, whereas inset representing the smallest overall bin shows bin sizes of 500 bp.

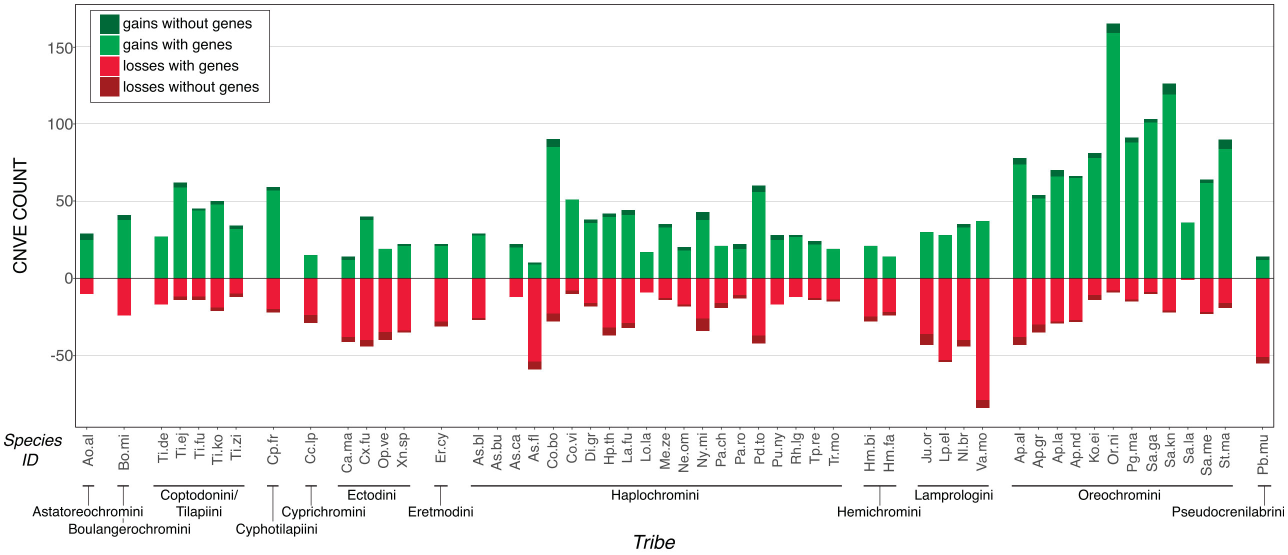

Fig. 5.—CNVE count per species, sorted by tribe. The majority of CNVE gains and losses in all species contain coding elements.

important to note the caveat that the number of species analyzed per tribe is highly variable with anywhere from 1 to 20 species per tribe.

Phylogenetics

To determine how species cluster according to CNVEs, we used RAxML to construct a tree based on our aCGH data (fig. 6). The CNVE tree does well in clustering both tribe and radiation designations with riverine species, which derive from diverse geographical locations, scattered among the clusters. The Haplochromini tribe is particularly well clustered by CNVE gains and losses. Specifically, the Lake Malawi Haplochromini share a set of CNVE gains that distinguish them from the Lake Victoria Haplochromini, which share a different set of gains as well as losses. Interestingly, the Lake

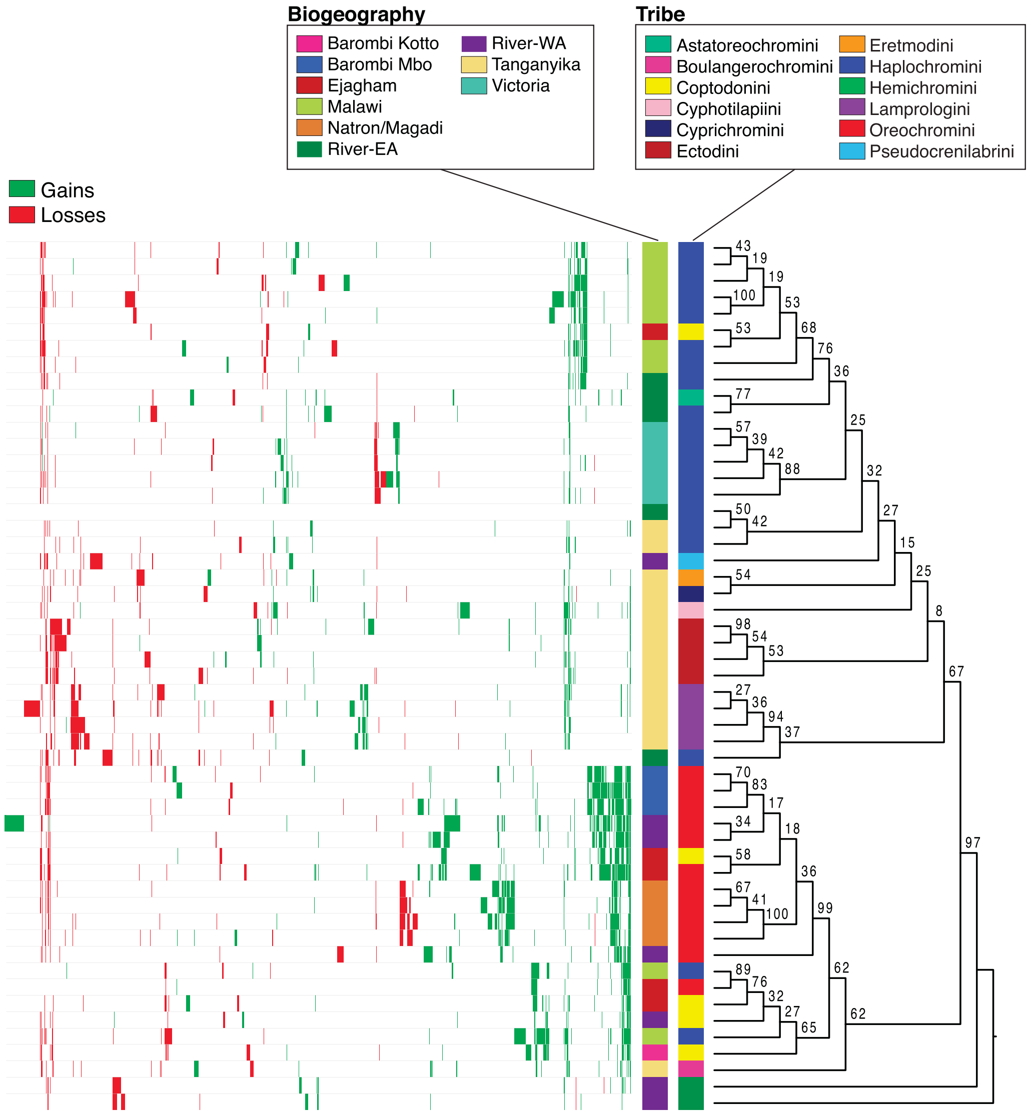

Fig. 6.—Heatmap showing RAxML clustering of CNVEs in all cichlid species using the model BINGAMMA. *Hemichromis fasciatus* (Hm.fa) was set as the outgroup, and bootstrap values are labeled at nodes. Despite low branch support at a majority of nodes, maximum likelihood tree corresponds well to tribe and radiation designations.

Victoria Haplochromini gains are also shared by several tribes from Lake Tanganyika. There are additional gains that appear in disparate Tanganyikan tribes including Cyphotilapini, Cyrichromini, Ectodini, and Eretmodini while excluding others such as the Lamprologini. Perhaps reflecting the fact that losses are less well tolerated than gains, we find fewer losses shared among species, however the Magadi species, and to a lesser extent the Lamprologini, do each host a set of largely unique losses.

To compare clustering patterns based on CNVE calls to phylogenetic relationships, we constructed a maximum likelihood tree based on *ND2* and *D-loop* mitochondrial sequences, which unsurprisingly, shows better correspondence to tribe designations. Nonetheless, much of the phylogenetic signal is recapitulated by the CNVE tree (fig. 7). TOPD analysis comparing the two topologies found 38 of the 52 species (Sa.kn is omitted for missing sequence data) placements disagree between trees and they share a split distance of 0.8163. Split distances range from 0 to 1, with values closer to 0 signaling that the two topologies are nearly identical. Although these statistics indicate the two trees differ in placement for a majority of taxa, CNVE and sequence-based cichlid phylogenies agree considerably better than in 100 random iterations of tree topologies for these data (average split distance: 0.99 ± 0.011 and disagreement: 51.13 ± 1.95/52 species). The observed disagreement between data sets is likely related in part to the low branch support in both trees (e.g., according to CNVEs As.fl is

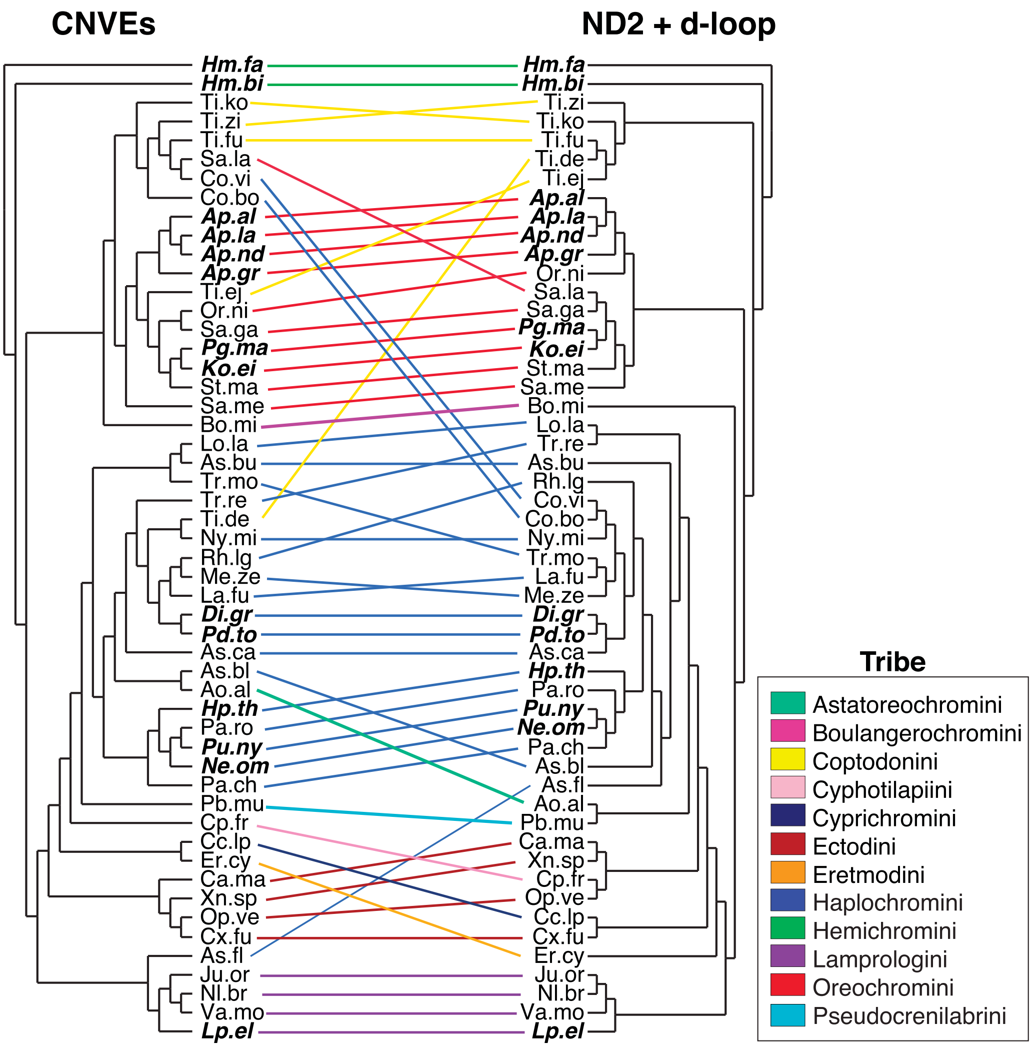

FIG. 7.—Comparison of RAxML trees using CNVE gain and loss data versus *ND2* and *d-loop* mitochondrial sequence data from 52/53 species examined in this study. RAxML models used were BINGAMMA and GTRGAMMA, respectively, for the two different data types. *Sarotherodon knauerae* (Sa.kn) was omitted for this comparison due to lack of quality sequence data for either of the two mitochondrial amplicons. Dendrograms are ordered for best alignment between data sets. Taxa highlighted in bold and italics are those that agree in exact placement between topologies as detected by TOPD.

inappropriately clustered with the Lamprologini at low bootstrap strength) (fig. 6). Even though exact placement of taxa may differ, these appear to be minor placement changes within consistent clade groupings (fig. 7).

Although the tree topologies are broadly similar and reflect uncontroversial cichlid phylogenetic relationships, there are some intriguing incongruencies between taxon placement in the mtDNA and CNVE-based trees (fig. 7). Although different evolutionary histories of nuclear and mitochondrial loci could contribute to this incongruence, it is well known that maternally inherited mitochondrial loci cannot reflect the entire history of the extant taxa where admixture has occurred. Furthermore, given that cichlid radiations have a history of interspecific hybridization (Joyce et al. 2011; Malinsky et al. 2018), incongruence between the mtDNA and CNVE trees is expected. Of the relationships that are most divergent between the two trees, two *Coptodon* species from Lake Ejagham, Ti.de (which shares many CNVE gains with other Haplochromini) and Ti.ej (which shares CNVE gains with other Oreochromini), have evidenced introgression (Martin et al. 2015). Similarly, our two *Copadichromis* species, Co.vi and Co.bo from Lake Malawi, have strong evidence for interspecific gene flow (Anseeuw et al. 2012), though the hybridizing species partner(s) remain unidentified. Such hybridization events likely impact CNVE gain/loss data and therefore the tree derived from that data. Independent (homoplasious) duplications of loci resulting from convergent evolution could also result in these incongruencies. Because gene duplication can be an important component of rapid adaptation, this might be an important window on how selection has shaped the gene complements of different cichlid groups. Full genome sequence for these CNVE regions in combination with more robust phylogenies will be necessary to resolve these possibilities. We look forward to a collection of much more taxonomically widespread genomes on which to test these hypotheses for the target loci identified here.

Genomic Architecture of CNVE Mapping

To understand the genomic architecture of these structural variations, we identify copy number hotspots using HD-CNV to indicate genomic regions with recurrent insertions and deletions among the 53 species (fig. 8 and supplementary fig. S2, Supplementary Material online). There are 51 detected copy number hotspots in cichlid genomes with 10 or more merged “CNV families” (considered replicate CNVs with 99% overlap, see Materials and Methods). The hotspot with the highest density of CNV families (17) is found on LG16-21 between ∼4.97 MB and ∼5.16 MB containing structural variation for 24 of 53 cichlid species across 4 tribes. According to gene annotations for the Or.ni reference genome, this hotspot contains at least four copies of *trace amine associated receptor 15* (*taar15*). TAARs are a family of G-protein-coupled receptors expressed in the olfactory epithelium (Liberles and Buck 2006; Hashiguchi and Nishida 2007) that are known to be highly copy number variable and colocalized in several teleost genomes (Chain et al. 2014; Gao et al. 2017). In cichlids, many copies have been localized to Or.ni LG16-21 (Azzouzi et al. 2015), although additional TAAR genes are found in hotspots on LG7. There are an additional 8 copy number hotspots on LG16-21, for a 77 total CNV families on this linkage group. Apart from LG16-21, only LG7 and LG14 have a greater number of CNV families, 93 and 82, and only 3 other linkage groups (LG22, LG8-24, and LG6) have more than 5 CNV hotspots composed of 10+ reciprocally overlapping CNV families. Nine linkage groups have no hotspots at this threshold including LG10 and LG9, which have the fewest CNV families with only 10 and 24, respectively. When we visualize these hotspots using the genome graphs tool in the University of California-Santa Cruz Genome Browser, we see that the majority of hotspots on each linkage group are clustered in one or two regions and are partially overlapping, suggesting that the 50% reciprocal overlap threshold is a conservative estimate of the level of recurrent copy number variation in a hotspot and may underestimate the level of structural dynamics in certain genomic regions (supplementary fig. S2, Supplementary Material online).

TE Analysis

To determine whether CNVEs are closely associated with specific TE families, we consider six broad categories of repetitive elements (DNA, LINE, LTR, RC, SINE, and Unknown) previously mapped in the Or.ni genome (Brawand et al. 2014). In most cases, the observed boundaries we detect for CNVEs are likely to be internal to the actual boundaries of CNVEs in the genome, which are expected fall between array probes. In order to target the actual CNVE boundaries, we also mapped TEs within 2- and 20-kb regions flanking predicted CNVEs. In both sample-level CNVEs and species-level CNVEs, we find DNA elements, LINEs, and LTRs to be significantly enriched within species-level CNVEs as well as with their 2- and 20-kb flanking regions (Bonferroni corrected $P < 0.01$). These elements are 25–175% more prevalent in CNVEs than in randomly selected matched loci, with LTRs showing the highest level of enrichment in both sets of CNVEs (fig. 9). Although rolling-circle (RC) transposons, aka helitrons, are rare in the Or.ni genome relative to other TE classes, we found they are also significantly enriched in both the defined sample-level and species-level CNVEs and the 20-kb flanking regions for species-level CNVEs (Bonferroni corrected $P < 0.05$, Bonferroni corrected $P < 0.05$). RC transposons specifically have been found to cause gene duplication and exon-shuffling in maize (Morgante et al. 2005), bats (Pritham and Feschotte 2007), and primates (Hedges and Batzer 2005), where they contribute to intraspecific structural variation. Overall our results implicate DNA elements, LINEs, LTRs, and RC transposons as major factors associated with CNVEs, possibly responsible for a large portion of copy number variation in cichlids and therefore important for the adaptive radiation of the clade. In Lepidoptera, many species show TE enrichment flanking CNVs however the specific TE type differs between species (Zhao et al. 2017). As our TE annotations are based directly on the Or.ni genome, a more nuanced analysis of TE load in the 53 species may reveal different enrichments, patterns, and conserved characteristics in the different clades. Furthermore, such analyses of TE insertions can be useful in reconciling phylogeny and inferring ancestry in adaptive radiations (Shedlock et al. 2004).

Gene Content

To determine the number and character of genes affected by CNVEs, we consider 30,385 unique features annotated for Or.ni in BROAD and Ensembl databases. Of the 1,413 species-level CNVEs, 1,300 contained a total of 7,404 annotated features representing 3,475 unique gene accessions. Within any single species, an average of 94.53% ± 2.03 of detected CNVEs contain at least one annotated feature suggesting that at least one gene was impacted (fig. 5). Although 800 CNVEs contain only a single annotated feature, one CNVE on unplaced scaffold UNK52 contains 80 features (supplementary fig. S3, Supplementary Material online), most of which are either immunoglobulin-related or novel genes. Gene-containing CNVEs are an average of ∼112.02 ± 189.11 kb in length, whereas CNVEs without genes are smaller at ∼52.55 ± 38.95 kb. The high rate of gene inclusion in a CNVE may be due to the gene-centric design of the array with greater probe density in annotated genes.

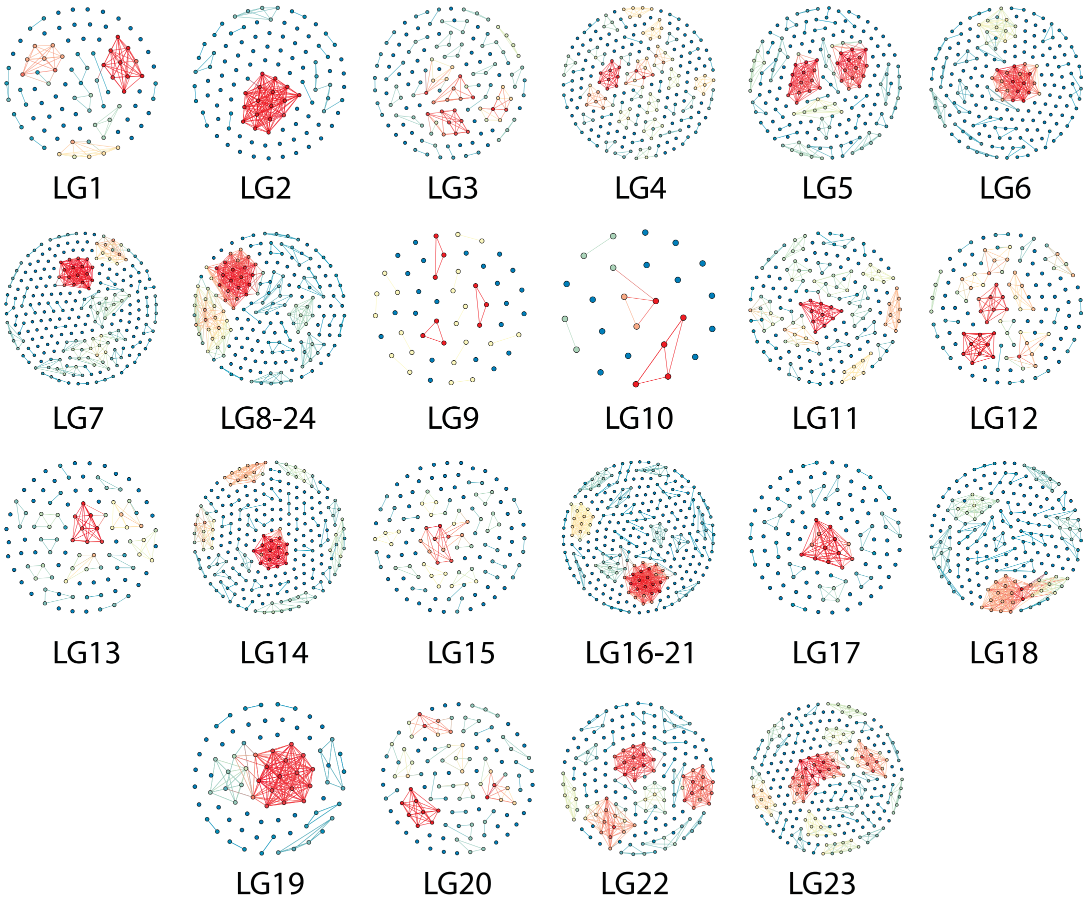

Fig. 8.—CNV hotspot map produced by HD-CNV. Input CNVs were concatenated outputs from DNAcopy segmentations from all individuals in study. Exact duplicate CNV coordinates were collapsed so all intervals were nonredundant, therefore this map is not biased toward recurrent called CNVs from As.bu reference samples. Nodes with warmer colors represent CNVs with higher numbers of unique overlapping CNVs and cool colors represent CNVs with fewer overlaps. HD-CNV parameters required 50% reciprocal overlap for CNV merges and 99% overlap for CNV families. Figure does not include any unplaced scaffolds.

When addressing functional impacts of copy number variation, we aimed to capture the full range of potentially impacted systems, therefore, we interrogate both the set of genes mapped within our 1,413 species-level CNVEs and those mapped within 2,879 sample-level CNVEs, which includes those found in a minority of samples for a given species demonstrating individual variation in cichlids. For both lists, we performed GO enrichment analysis (fig. 10 and supplementary tables S4 and S5, Supplementary Material online) and found nine GO categories enriched in both gene lists largely relating to olfactory sensing and ubiquitination, three exclusive to the species-level CNVEs ("endonuclease activity," "proteolysis," and "apical junction complex"), and eight exclusive to the sample-level list with its additional variation ("antigen processing and presentation of peptide antigen via MHC class I," "protein binding," "GTP binding," "receptor-mediated endocytosis," "immune response," "solute:hydrogen antiporter activity," "peptide antigen binding," and "MHC class I protein complex"). The latter set of exclusive enriched GO terms contains many immune-related categories, suggesting there is a substantial amount of copy number variation of immune genes within species.

Although previous analysis of gene duplicates identified from cichlid genome assemblies (Brawand et al. 2014) noted opsins as the only functional category of genes enriched among CNVs, our aCHG analysis more closely parallels early aCGH results (Machado et al. 2014) and identifies GO terms that have previously been associated with adaptation to diverse environments and reflect gene categories noted to proliferate. For example, immune system genes and proteases are known to evolve rapidly following duplication. We find the term "antigen processing and presentation of peptide antigen via MHC class 1" to be enriched. The MHC class I molecules allow each cell to provide a readout of protein expression to be monitored by cytotoxic T lymphocytes and natural killer

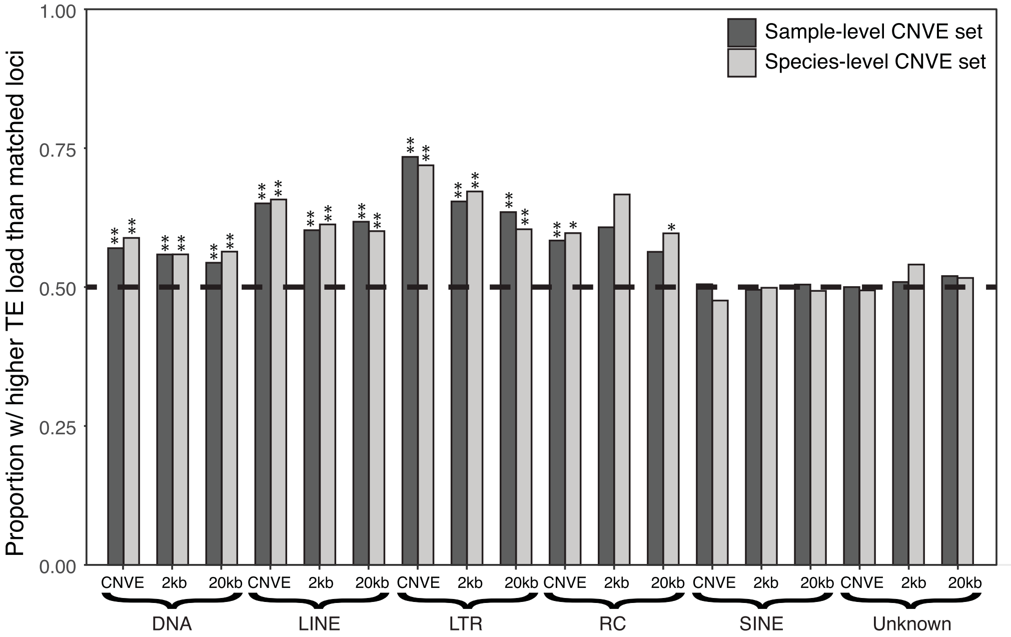

Fig. 9.—Enrichment of six classes of repetitive elements in sample-level and species-level CNVEs and 5′/3′ flanking regions as determined by binomial tests. CNVE 2- and 20-kb flanking regions were tested to capture actual region boundaries, accounting for underestimate of actual CNVE length due to array probe spacing in Or.ni genome. Results are presented as the proportion of observed CNVEs that contain more TEs of each class than randomly selected genomic intervals matched for approximate length, probe number, and sequence of probe types (exonic vs. noncoding). *$P <$ Bonferroni corrected 0.05. **$P <$ Bonferroni corrected 0.01.

cells. The polymorphic alleles which confer differential susceptibilities to infection have evolved in response, in part, to viruses that have evolved mechanisms to hijack this pathway (Hewitt 2003). Another term, "serine-type endopeptidase activity" reflects genes involved in many physiological functions that could play a role in adaptation to new environments such as digestion, immune response, blood coagulation, and reproduction (Di Cera 2009). These are among the families of proteases that have undergone dramatic expansion in the metazoans and have been found to be enriched among segmental duplication in a study of Lepidoptera (Zhao et al. 2017). Similarly, the term "Scavenger receptors" includes extracellular glycoprotein receptors important for the removal of waste materials and foreign substances including bacteria and can play an important role in adaptation to novel environments (Yap et al. 2015). The term "ubiquitin-protein ligase activity" applies to genes that play an important role in substrate specificity, the ubiquitination pathway, regulation of cell trafficking, DNA repair, and signaling (Glessner et al. 2009). Among the other GO terms we find enriched, "Olfactory receptor activity" and several "G-protein-coupled receptor" terms represent a group of genes known for their rapid expansion and diversification related to the detection of chemical stimulus.

Five of the enriched GO terms are also enriched specifically in copy number hotspots detected with HD-CNV (fig. 10 and supplementary table S6, Supplementary Material online). "G-protein-coupled receptor activity," "G-protein-coupled receptor signaling pathway," "acyl-CoA metabolic process," "detection of chemical stimulus involved in sensory perception of smell," and "olfactory receptor activity" are all enriched in highly recurrent copy number hotspots. Nine detected hotspots spread over five linkage groups (LG7, LG8-24, LG11, LG16-21, and LG17) are found in four or more cichlid tribes, and this subset is enriched for "G-protein-coupled receptor activity" and "G-protein-coupled receptor signaling pathway," GO terms. These enriched categories represent the most copy number variable genomic regions across the cichlid phylogeny, and they closely mirror those found in copy number variable regions in stickleback (Feulner et al. 2013; Chain et al. 2014), another clade known for its propensity to speciate. Taken together these results suggest that the remarkable individual variation that resides within our identified hotspots represents variation for genes with roles in adaptive phenotypes that could reasonably promote divergence and speciation in a clade.

When we examine GO term enrichment by tribe, five different tribes of cichlids produced no enriched GO terms. Among the seven tribes with significant GO terms, they are nearly all enriched for the four categories associated with the detection of and response to environmental chemical cues ("G-protein-coupled receptor signaling pathway," "G-protein-coupled receptor activity," "Olfactory receptor activity," and "detection of chemical stimulus involved in sensory

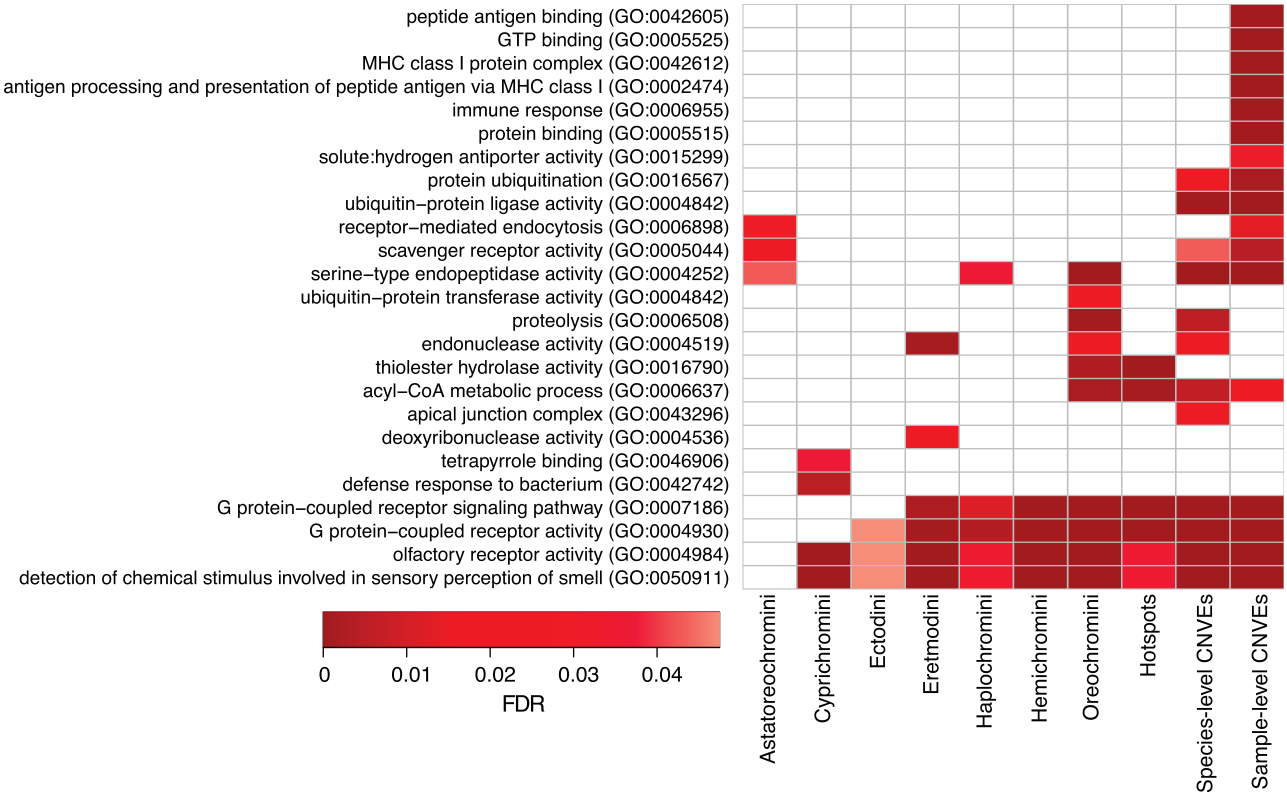

Fig. 10.—FDR heatmap for enriched GO categories of genes within subsets of observed CNVEs. Subsets include sample-level CNVEs, species-level CNVEs, CNVE hotspots, and CNVEs represented within each tribe. Tribes not listed in figure had no enriched GO categories. Each enrichment test set contains genes overlapping CNVEs and the reference set is the entire set of genes in the annotated Or.ni genome. Blank cells are not significant at FDR < 0.05.

perception of smell") (fig. 10). This adds to a growing body of evidence that genes underlying olfaction are highly divergent in both copy number and type in cichlids and many other taxa (Nei et al. 2008; Young et al. 2008; Brawand et al. 2014). Olfactory sensing is thought to be important for sexual and natural selection and as a result may contribute to reproductive isolation and speciation (Salzburger 2009), which likely explains the extensive copy number variation we identify between species. Aside from GO terms related to olfaction, "serine-type endopeptidase activity" is enriched among three tribes and "endonuclease activity" is enriched in two tribes. All other significant GO terms are specific to a single tribe. Interestingly, four significant GO terms specific to single tribes are not enriched when looking at the overall species-level or sample-level CNVE gene sets, including "ubiquitin-protein transferase activity" in Oreochromini, "deoxyribonuclease activity" in Eretmodini, and "tetrapyrrole binding" and "defense response to bacterium" in Cyprichromini. Although follow-up analysis is required, these categories may point to specific adaptive sets of genes for particular cichlid clades. For example, "defense response to bacterium" is identified as significant in Cyprichromini due to an exclusive duplication in *Cyprichromis leptosoma* (Cc.lp) on the contig UNK44 containing moronecidin and several *moronecidin-like* antimicrobial genes. This suite of CNV genes may protect Cc.lp from novel pathogens present in its niche habitat (Karvonen et al 2018). In general, fish encounter a wide range of pathogenic microorganisms, thus the innate immune system represents an important potential axis for adaptation that has been investigated in cichlids. One player, the *c-type lysozyme* gene, that has been shown to be duplicated in some cichlid species (*O. niloticus*, *L. caeruleus* redtail sheller, and possibly *C. leptosome*) (Takahashi-Kariyazono et al. 2017), resides within a large CNVE which shows gains in some samples from several species (Or.ni, Ko.ei, Sa.la, Co.vi, Ao.al, and Ti.zi). Though the species analyzed in both studies have little overlap, these finding corroborate known CNVs for genes of adaptive function.

In addition to the opsins and *c-type lysozyme* genes discussed above, for which gene duplication is known to have played a role in adaptations to both the abiotic and biotic environment including social interactions, several sequence variants of other genes located within CNVEs have been previously studied in relation to key evolutionary adaptations in cichlids. For example, hemoglobin subunits (represented in our CNVE set by *hemoglobin subunit alpha-D*) are thought to constitute a "supergene" (Hahn et al. 2017), which is known to show signatures of selection correlated with habitat depth (Malinsky et al 2018) and anthropogenic disturbance (Witte et al. 2013). With regard to color pattern diversity, we find *agouti-related peptide 2*, which harbors mutations recently shown to underlie the convergent evolution of horizontal melanic stripe patterns in lineages from Lake Malawi, Victoria, and Tanganyika (Kratochwil et al. 2018). However, the two rapidly evolving pigmentation gene paralogs *fhl2a* and *fhlb* that are known to play a role in egg-spot formation (Santos et al. 2014) important for mating behavior and species isolation are not found within CNVE regions. With regard to morphological diversity, we identify several members of the wnt signaling pathway (*wnt2*, *wnt7bb*, *wisp2*, *wnt7ab*, and *wnt4*) contained within the CNVEs. This pathway has been implicated in the evolution of impressive craniofacial diversity (Parsons et al. 2014; Powder et al. 2015) as well as body and fin morphology (Navon et al. 2017). However, another well-known player in craniofacial diversity, *bone morphogenetic protein 4* (*bmp4*) (Streelman and Albertson 2006), does not reside within a CNVE. Overall, these data suggest that individual genes important for physiological, behavioral, and morphological adaptations may underlie diversity through sequence variation, copy number variation (as identified in this data set), or both.

Conclusions

Here, we describe the genomic diversity with regard to variation in DNA copy number found among the African cichlid assemblage that represents morphological, ecological, and behavioral diversity. We demonstrate that gene duplication, which has the potential to generate substantial molecular substrate for the origin of evolutionary novelty, can be assayed through aCGH. The techniques applied here do not reveal whether the duplication originated through transposition, retrotransposition (Brosius 1991), segmental duplication (Bailey et al. 2001), tandem duplication (Katju and Lynch 2003; Nozawa and Nei 2007), or change in chromosomal or genomic ploidy (Van de Peer et al. 2009; Sato and Nishida 2010), although CNV-rich regions are found to also be enriched for certain TEs. Nor does it address whether silencing, dosage compensation, or neo- or sub-functionalization (Lynch and Conery 2000) constitute the fate or functional retention of these duplicated regions. The aCGH technique is most effective for the discovery of highly similar gene duplicates (i.e., evolutionarily recent or highly conserved) and thus complements sequence-based approaches in which these events are likely to be collapsed in the assembly of short-read sequence data. By performing a replicate analysis for the species previously analyzed by read-depth techniques (Brawand et al. 2014), spotted cDNA array (Machado et al. 2014), and a new genome assembly based on long-read sequence data (Conte et al. 2017), we are able to validate our pipeline and present results consistent with these empirical studies as well as theoretical hypotheses (Seehausen 2006) (supplementary information, Supplementary Material online).

By addressing intraspecific variation for Me.ze, Nl.br, and Pu.ny, we discovered an average of only 50–100 CNVEs per individual and 150–200 total CNVEs within a species, such that only 10% of the detected CNVEs appear fixed for a species. This high level of detected intraspecific variation suggests that many of the reported species-level CNVEs analyzed in species with fewer samples, although real, may instead represent sample-level variation. This intraspecific variation provides substrate for adaptation and evolution.

Despite the inclusion of CNVEs that may represent intraspecific variation, our RAxML tree based on CNVEs does cluster species well at both the tribe and the radiation level, approximately recapitulating inferred phylogenetic relationships. Although some species (e.g., the tribe Ectodini) actually cluster better according to the CNVE-based tree, those that do not (e.g., *Copadichromis* and *Tilapia* species) suggest the hypothesis that these clades may be experiencing a greater rate of structural rearrangements. In part, the mismatch between the CNVE-based tree and mitochondrial-based phylogenetic tree may be mediated through active TEs, considering CNVE regions are highly enriched for certain types of TEs identified in the Or.ni genome, particularly DNA elements, LINEs, LTRs, and helitrons. Although the gene-centric design of our array may bias the results somewhat, we note that the vast majority of the detected CNVEs (∼95%) contain at least one annotated feature and GO analysis suggests that these are enriched for genes belonging to functional categories with potentially adaptive phenotypic consequences. Therefore, genomes with TEs near these categories of genes may predispose a lineage to radiation.

A complete understanding of the molecular basis for adaptive natural selection, speciation, and even adaptive radiation requires further study of copy number variation. Taken together, our results reveal not only a high level of individual variation but also substantial repeated evolution resulting in hotspots, many of which show enrichment for genes from functional categories that suggest potentially adaptive roles. As such, the reported CNVEs are likely to have played a role in the divergence and speciation observed among cichlids.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We are thankful to all the cichlid researchers who contributed samples, as well as Shaohua Fan who contributed the read-depth data set. Reed students, Jaclyn Calkins, Ginger Jui, Karl Menzel, and Jeff Hunter contributed to this project at early stages. The authors confirm that there are no known conflicts of interest associated with this publication. This work was supported by NSF-DEB 1021582 to S.C.P.R.

Literature Cited

Altschul SF, et al. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Anseeuw D, et al. 2012. Extensive introgression among ancestral mtDNA lineages: phylogenetic relationships of the Utaka within the Lake Malawi cichlid flock. Int J Evol Biol. 2012:865603.

Azzouzi N, et al. 2015. Identification and characterization of cichlid TAAR genes and comparison with other teleost TAAR repertoires. BMC Genomics. 16(1):335.

Bailey JA, et al. 2001. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 11(6):1005–1017.

Baldwin BG, Sanderson MJ. 1998. Age and rate of diversification of the Hawaiian Silversword alliance (*Compositae*). Proc Natl Acad Sci U S A. 95(16):9402–9406.

Bastian M, et al. 2009. Gephi: An open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media.

Bloomfield G, et al. 2008. Widespread duplications in the genomes of laboratory stocks of *Dictyostelium discoideum*. Genome Biol. 9(4):R75.

Braasch I, et al. 2006. Asymmetric evolution in two fish-specifically duplicated receptor tyrosine kinase paralogons involved in teleost coloration. Mol Biol Evol. 23(6):1192–1202.

Brawand D, et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. Nature 513(7518):375–381.

Brosius J. 1991. Retroposons—seeds of evolution. Science 251(4995):753.

Brunelle BW, et al. 2004. Microarray-based genomic surveying of gene polymorphisms in *Chlamydia trachomatis*. Genome Biol. 5(6):R42.

Butler JL, et al. 2013. HD-CNV: hotspot detector for copy number variants. Bioinformatics 29(2):262–263.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Carleton KL, Kocher TD. 2001. Cone opsin genes of African cichlid fishes: tuning spectral sensitivity by differential gene expression. Mol Biol Evol. 18(8):1540–1550.

Chain FJJ, et al. 2014. Extensive copy-number variation of young genes across stickleback populations. PLoS Genet. 10(12):e1004830.

Chen CC, Fernald RD. 2006. Distributions of two gonadotropin-releasing hormone receptor types in a cichlid fish suggest functional specialization. J Comp Neurol. 495(3):314–323.

Chen Z, et al. 2008. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc Natl Acad Sci U S A. 105(35):12944–12949.

Cnaani A, Kocher TD. 2008. Sex-linked markers and microsatellite locus duplication in the cichlid species *Oreochromis tanganicae*. Biol Lett. 4(6):700–703.

Conte MA, Kocher T. 2015. An improved genome reference for the African cichlid, *Metriaclima zebra*. BMC Genomics. 16:724.

Conte MA, et al. 2017. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. BMC Genomics. 18(1):341.

Coppe A, et al. 2013. Genome evolution in the cold: Antarctic icefish muscle transcriptome reveals selective duplications increasing mitochondrial function. Genome Biol Evol. 5(1):45–60.

Darwin C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: Watts & Co.

Di Cera E. 2009. Serine proteases. IUBMB Life 61(5):510–515.

Dixon WJ. 1950. Analysis of extreme values. Ann Math Statist. 21(4):488–506.

Dobzhansky T. 1937. Genetics and the origin of species. 2nd ed., 1941; 3rd ed., 1951. New York: Columbia University Press.

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proc Natl Acad Sci U S A. 104(50):19920–19925.

Feulner PGD, et al. 2013. Genome-wide patterns of standing genetic variation in a marine population of three-spined sticklebacks. Mol Ecol. 22(3):635–649.

Fryer G. 1991. Comparative aspects of adaptive radiation and speciation in Lake Baikal and the great rift lakes of Africa. Hydrobiologia 211(2):137–146.

Fryer G, Iles TD. 1972. The cichlid fishes of the great lakes of Africa: their biology and evolution. Edinburgh (United Kingdom): Oliver & Boyd.

Gao S, et al. 2017. Genomic organization and evolution of olfactory receptors and trace amine-associated receptors in channel catfish, *Ictalurus punctatus*. Biochim Biophys Acta Gen Subj. 1861(3):644–651.

Gazave E, et al. 2011. Copy number variation analysis in the great apes reveals species-specific patterns of structural variation. Genome Res. 21(10):1626–1639.

Gilbert LB, et al. 2011. Array CGH phylogeny: how accurate are comparative genomic hybridization-based trees? BMC Genomics. 12:487.

Glessner JT, et al. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459(7246):569–573.

Gotz S, et al. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420–3435.

Hahn C, et al. 2017. The genomic basis of cichlid fish adaptation within the deepwater "twilight zone" of Lake Malawi. Evol Lett. 1(4):184–198.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered. 100(5):605–617.

Han MV, et al. 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res. 19(5):859–867.

Hashiguchi Y, Nishida M. 2007. Evolution of trace amine associated receptor (TAAR) gene family in vertebrates: lineage-specific expansions and degradations of a second class of vertebrate chemosensory receptors expressed in the olfactory epithelium. Mol Biol Evol. 24(9):2099–2107.

Hedges DJ, Batzer MA. 2005. From the margins of the genome: mobile elements shape primate evolution. BioEssays 27(8):785–794.

Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes. Genome Res. 17(12):1837–1849.

Hewitt EW. 2003. The MHC class I antigen presentation pathway: strategies for viral immune evasion. Immunology 110(2):163–169.

Hulsey CD, et al. 2018. Phylogenomics of a putatively convergent novelty: did hypertrophied lips evolve once or repeatedly in Lake Malawi cichlid fishes? BMC Evol Biol. 18:179.

Huson DH, et al. 2007. Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460.

Ivory SJ, et al. 2016. Environmental change explains cichlid adaptive radiation at Lake Malawi over the past 1.2 million years. Proc Natl Acad Sci U S A. 113(42):11895–11900.

Joyce DA, et al. 2011. Repeated colonization and hybridization in Lake Malawi cichlids. Curr Biol. 21(3):R108–R109.

Karvonen A, et al. 2018. Divergent parasite infections in sympatric cichlid species in Lake Victoria. J Evol Biol. 31(9):1313–1329.

Kassen R. 2009. Toward a general theory of adaptive radiation. Insights from microbial experimental evolution. Ann N Y Acad Sci. 1168:3–22.

Katju V, Lynch M. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. Genetics 165(4):1793–1803.

Kearse M, et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics (Oxford, Engl). 28(12):1647–1649.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci. 279(1749):5048–5057.

Kratochwil CF, et al. 2018. Cis-regulatory changes in agrp2 impact horizontal melanic stripes in multiple species thus representing convergent evolution of a complex color pattern. Science 362(6413):457–460.

Kück P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. Mol Phylogenet Evol. 56(3):1115–1118.

Liberles SD, Buck LB. 2006. A second class of chemosensory receptors in the olfactory epithelium. Nature 442(7103):645–650.

Locke DP, et al. 2003. Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res. 13(3):347–357.

Losos JB, et al. 1998. Contingency and determinism in replicated adaptive radiations of island lizards. Science 279(5359):2115–2118.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290(5494):1151–1155.

Lynch M, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A. 105(27):9272–9277.

Machado HE, et al. 2010. Genomic architecture of adaptive radiation: the role for gene duplication in African cichlid fishes. Integr Comp Biol. 50:E262.

Machado HE, et al. 2014. Gene duplication in an African cichlid adaptive radiation. BMC Genomics. 15(1):161.

Malinsky M, et al. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat Ecol Evol. 2(12):1940–1955.

Martin CH, et al. 2015. Complex histories of repeated gene flow in Cameroon crater lake cichlids cast doubt on one of the clearest examples of sympatric speciation. Evolution 69(6):1406–1422.

Mayr E. 1963. Animal species and evolution. Cambridge (MA): The Belknap Press of Harvard University Press.

Meyer BS, Matschiner M, Salzburger W. 2016. Disentangling incomplete lineage sorting and introgression to refine species-tree estimates for Lake Tanganyika cichlid fishes. Syst Biol. 66:531–550.

Morgante M, et al. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 37(9):997–1002.

Navon D, et al. 2017. Genetic and developmental basis for fin shape variation in African cichlid fishes. Mol Ecol. 26(1):291–303.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet. 9(12):951–963.

Neph S, et al. 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28(14):1919–1920.

Nozawa M, Nei M. 2007. Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. Proc Natl Acad Sci U S A. 104(17):7122–7127.

Ohno S. 1970. Evolution by gene duplication. Berlin, Heidelberg (Germany): Springer.

Panova M, et al. 2014. Species and gene divergence in *Littorina* snails detected by array comparative genomic hybridization. BMC Genomics. 15:687.

Parsons KJ, et al. 2014. Wnt signaling underlies the evolution of new phenotypes and craniofacial variability in Lake Malawi cichlids. Nat Commun. 5:3629.

Powder KE, Milch K, Asselin G, Albertson RC. 2015. Constraint and diversification of developmental trajectories in cichlid facial morphologies. Evodevo 6:25.

Pritham EJ, Feschotte C. 2007. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. Proc Natl Acad Sci U S A. 104(6):1895–1900.

Puigbò P, Garcia-Vallvé S, McInerney JO. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. Bioinformatics 23(12):1556–1558.

Quinlan AR, Hall IM. 2010. BEDtools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.

Redon R, et al. 2006. Global variation in copy number in the human genome. Nature 444(7118):444–454.

Renn SCP, et al. 2010. Using comparative genomic hybridization to survey genomic sequence divergence across species: a proof-of-concept from *Drosophila*. BMC Genomics. 11(1):271.

Riehle MM, et al. 2006. Natural malaria infection in *Anopheles gambiae* is regulated by a single genomic control region. Science 312(5773):577–579.

Ritchie ME, et al. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43(7):e47.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53(1-2):131–147.

RStudio Team. 2015. RStudio: integrated development for R. Boston (MA): RStudio, Inc. Available from: http://www.rstudio.com/, last accessed August 27, 2019.

Santini S, Bernardi G. 2005. Organization and base composition of tilapia hox genes: implications for the evolution of hox clusters in fish. Gene 346:51–61.

Salzburger W. 2009. The interaction of sexually and naturally selected traits in the adaptive radiations of cichlid fishes. Mol Ecol. 18(2):169–185.

Salzburger W. 2018. Understanding explosive diversification through cichlid fish genomics. Nat Rev Genet. 19(11):705–717.

Santos EM, et al. 2014. The evolution of cichlid fish egg-spots is linked with a cis-regulatory change. Nat Commun. 5:5149.

Sato Y, Nishida M. 2010. Teleost fish with specific genome duplication as unique models of vertebrate evolution. Environ Biol Fish. 88(2):169–188.

Schluter D. 2000. The ecology of adaptive radiation. Oxford: Oxford University Press.

Seehausen O. 2006. African cichlid fish: a model system in adaptive radiation research. *Proceedings of the Royal Society of London B: Biological Sciences*. 273(1597): 1987–1998.

Seehausen O, et al. 2008. Speciation through sensory drive in cichlid fish. Nature 455(7213):620–626.

Seshan VE, Olshen A. 2016. DNAcopy: DNA copy number data analysis. R package version 1.34.0.

Shedlock AM, et al. 2004. SINEs of speciation: tracking lineages with retroposons. Trends Ecol Evol (Amst). 19(10):545–553.

Shirak A, et al. 2008. Copy number variation of lipocalin family genes for male-specific proteins in tilapia and its association with gender. Heredity (Edinb). 101(5):405–415.

Skinner BM, et al. 2014. Global patterns of apparent copy number variation in birds revealed by cross-species comparative genomic hybridization. Chromosome Res. 22(1):59–70.

Spady TC, et al. 2005. Adaptive molecular evolution in the opsin genes of rapidly speciating cichlid species. Mol Biol Evol. 22(6):1412–1422.

Spady TC, et al. 2006. Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. Mol Biol Evol. 23(8):1538–1547.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Streelman JT, Albertson RC. 2006. Evolution of novelty in the cichlid dentition. J Exp Zool B Mol Dev Evol. 306(3):216–226.

Sudmant PH, et al. 2015. An integrated map of structural variation in 2, 504 human genomes. Nature 526(7571):75–81.

Sugie A, et al. 2004. The evolution of genes for pigmentation in African cichlid fishes. Gene 343(2):337–346.

Summers K, Zhu Y. 2008. Positive selection on a prolactin paralog following gene duplication in cichlids: adaptive evolution in the context of parental care. Copeia 2008(4):872–876.

Takahashi-Kariyazono S, et al. 2017. Gene duplications and the evolution of c-type lysozyme during adaptive radiation of East African cichlid fish. Hydrobiologia 791(1):7–20.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet. 38:615–643.

Terai Y, et al. 2006. Divergent selection on opsins drives incipient speciation in Lake Victoria cichlids. PLoS Biol. 4(12):e433.

Toedling J, et al. 2007. Ringo—an R/Bioconductor package for analyzing chip-chip readouts. BMC Bioinformatics 8:221.

Turner GF. 2007. Adaptive radiation of cichlid fish. Curr Biol. 17(19):R827–R831.

Turner TL, et al. 2005. Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol. 3(9):e285.

Van de Peer Y, et al. 2009. The flowering world: a tale of duplications. Trends Plant Sci. 14(12):680–688.

Wagner CE, et al. 2012. Ecological opportunity and sexual selection together predict adaptive radiation. Nature 487(7407):366–370.

Ward JH. 1963. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 58(301):236–244.

Warnes GR, et al. 2016. GPlots: various R programming tools for plotting data. R package version 3.0.1. The Comprehensive R Archive Network https://CRAN.R-project.org/package=gplots.

Watanabe M, et al. 2007. Functional diversification of kir7.1 in cichlids accelerated by gene duplication. Gene 399(1):46–52.

Witte F, et al. 2013. Cichlid species diversity in naturally and anthropogenically turbid habitats of Lake Victoria, East Africa. Aquat Sci. 75(2):169–183.

Yap NVL, et al. 2015. The evolution of the scavenger receptor cysteine-rich domain of the class a scavenger receptors. Front Immunol. 6:342.

Young JM, et al. 2008. Extensive copy-number variation of the human olfactory receptor gene family. Am J Hum Genet. 83(2):228–242.

Yu P, et al. 2013. Genome-wide copy number variations in *Oryza sativa* l. BMC Genomics. 14:649.

Zhang JZ. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18(6):292–298.

Zhao Q, et al. 2017. Segmental duplications: evolution and impact among the current Lepidoptera genomes. BMC Evol Biol. 17(1):161.

Associate editor: B. Venkatesh