Lake Malawi cichlid pangenome graph reveals extensive structural variation driven by transposable elements

Fu Xiang Quah, Miguel Vasconcelos Almeida, Moritz Blumer, et al.

Genome Res. 2025 35: 1094-1107 originally published online April 10, 2025
Access the most recent version at doi:10.1101/gr.279674.124

Research

Lake Malawi cichlid pangenome graph reveals extensive structural variation driven by transposable elements

Fu Xiang Quah,[1,2] Miguel Vasconcelos Almeida,[1] Moritz Blumer,[2] Chengwei Ulrika Yuan,[1,2] Bettina Fischer,[2] Kirsten See,[1] Ben Jackson,[2] Richard Zatha,[3] Bosco Rusuwa,[3] George F. Turner,[4] M. Emília Santos,[5] Hannes Svardal,[6] Martin Hemberg,[7] Richard Durbin,[2] and Eric Miska[1,2]

[1]*Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, United Kingdom;* [2]*Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom;* [3]*Department of Biological Sciences, University of Malawi, P.O. Box 280, Zomba, Malawi;* [4]*School of Environmental and Natural Sciences, Bangor University, Bangor, Gwynedd LL57 2TH, United Kingdom;* [5]*Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, United Kingdom;* [6]*Department of Biology, University of Antwerp, 2610 Wilrijk, Belgium;* [7]*The Gene Lay Institute of Immunology and Inflammation, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA*

Pangenome methods have the potential to uncover hitherto undiscovered sequences missing from established reference genomes, making them useful to study evolutionary and speciation processes in diverse organisms. The cichlid fishes of the East African Rift Lakes represent one of nature's most phenotypically diverse vertebrate radiations, but single-nucleotide polymorphism (SNP)–based studies have revealed little sequence difference, with 0.1%–0.25% pairwise divergence between Lake Malawi species. These were based on aligning short reads to a single linear reference genome and ignored the contribution of larger-scale structural variants (SVs). We constructed a pangenome graph that integrates six new and two existing long-read genome assemblies of Lake Malawi haplochromine cichlids. This graph intuitively represents complex and nested variation between the genomes and reveals that the SV landscape is dominated by large insertions, many exclusive to individual assemblies. The graph incorporates a substantial amount of extra sequence across seven species, the total size of which is 33.1% longer than that of a single cichlid genome. Approximately 4.73% to 9.86% of the assembly lengths are estimated as interspecies structural variation between cichlids, suggesting substantial genomic diversity underappreciated in SNP studies. Although coding regions remain highly conserved, our analysis uncovers a significant proportion of SV sequences as transposable element (TE) insertions, especially DNA, LINE, and LTR TEs. These findings underscore that the cichlid genome is shaped both by small-nucleotide mutations and large, TE-derived sequence alterations, both of which merit study to understand their interplay in cichlid evolution.

[Supplemental material is available for this article.]

The promise of pangenome methods lies in their ability to provide a more holistic view of genomic diversity by avoiding the constraints of a single reference genome (Eizenga et al. 2020). Originally developed for prokaryotes (Tettelin et al. 2005), the idea of conceptualizing the genetic repertoire in bacterial strains as core and dispensable genes has contributed to our understanding of pathogenicity and horizontal gene transfer (Rasko et al. 2008). Over the past decade, increasingly affordable long-read technologies (Eid et al. 2009; Mikheyev and Tin 2014) have allowed the adaptation of these methods to eukaryotic systems. Pangenomic studies in these organisms generally focus on quantifying differences at the nucleotide level, owing to the more stable nature of eukaryotic gene content (Vernikos et al. 2015; Sherman and Salzberg 2020). These pangenomes are frequently modeled as genome graphs (Garrison et al. 2018; Li et al. 2020), which work well to represent and detect structural variants (SVs), sequence alterations >50 bp, compared with reference-based short-read alignment (Mérot et al. 2020). Pangenomic approaches have proven valuable in human studies to uncover genomic diversity in underrepresented populations (Sherman et al. 2019; Gao et al. 2023; Liao et al. 2023) but have also been used to reveal novel sequences associated with complex traits in livestock and agriculturally relevant plants (Gong et al. 2023), such as pathogen immunity in cattle (Crysnanto et al. 2021), body development in ducks (Wang et al. 2024) and sheep (Li et al. 2023), fruit flavor in tomatoes (Gao et al. 2019), and flowering time in cucumbers (Li et al. 2022). In wild populations, the use of pangenomes is gaining attention, as they can uncover key genetic features relevant to evolutionary processes that are missing in established reference genomes (Secomandi et al. 2023; Fang and Edwards 2024; Plessy et al. 2024), making them powerful for studying species complexes and understanding diverse organisms.

Corresponding authors: fxq20@cam.ac.uk, eam29@cam.ac.uk
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279674.124. Freely available online through the *Genome Research* Open Access option.

35:1094–1107 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/25; www.genome.org

The cichlid fishes in the East African Rift Lakes are among nature's most spectacular vertebrate radiations, forming excellent systems to study evolution and speciation (Fryer and Iles 1972; Salzburger 2018; Santos et al. 2023). However, the contribution of large-scale SVs to cichlid adaptation is poorly understood, with most population studies having primarily focused on single-nucleotide polymorphisms (SNPs) detected by reference-based short-read alignment. Although the Victoria, Malawi, and Tanganyika lakes each host hundreds of phenotypically diverse species with an exuberance of sizes, morphologies, colors, diets, and behaviors (Schulte et al. 2014; Albertson et al. 2018; York et al. 2018; Hulsey et al. 2020), there are extraordinarily low rates of genetic divergence within and between species as measured by SNPs (Svardal et al. 2020). In particular, pairwise SNP divergence among Lake Malawi species stood at a mere 0.1%–0.25% (Malinsky et al. 2018), approximately 10 times lower than that between humans and chimpanzees (The Chimpanzee Sequencing and Analysis Consortium 2005). These genome-wide studies generally ignore SVs owing to the inherent difficulty in detecting them, although there are known examples of their regulatory roles in cichlid biology, including in vision (Schulte et al. 2014; Carleton et al. 2020; Nandamuri et al. 2023), sex determination (Munby et al. 2021), body coloring (Kratochwil et al. 2022), and egg spot patterning (Santos et al. 2014), many of which are attributable to transposable element (TE) insertions. An early attempt to study SVs in cichlids involved syntenic comparisons of the Lake Malawi species *Maylandia zebra* with the much more widely distributed cichlid *Oreochromis niloticus* (Conte et al. 2019). Another study analyzed SVs across the wider East African radiation (Penso-Dolfin et al. 2020) but utilized highly fragmented Illumina-based assemblies (Brawand et al. 2014), limiting its ability to detect more complex genomic variation.

There is currently limited insight into SVs in cichlids via a genome-wide approach, especially within individual radiations like Lake Malawi, which harbors an estimated more than 800 species predominantly from the Haplochromini tribe (Konings 1989). In this study, we constructed a pangenome graph to study Lake Malawi cichlid diversity, integrating six newly sequenced and two previously published long-read assemblies of select haplochromine species, spanning most of the major ecomorphological groups in the lake. We investigate the utility of the pangenome graph to discover SVs and identify polymorphic TE insertions, complementing conventional SNP-based techniques.

Results

Long-read assemblies and graph construction

SNP studies have provided support for the categorization of Lake Malawi cichlids into seven ecomorphological groups (Fig. 1A; Malinsky et al. 2018): (1) the littoral, rock dwelling "mbuna"; (2) the river and swamp-dwelling *Astatotilapia calliptera*; (3) benthics in shallow areas; (4) benthics in deep areas; (5) the "utaka"; and the pelagic groups (6) *Diplotaxodon* and (7) *Rhamphochromis*. For this study, six long-read assemblies were newly prepared for five species, specifically one mbuna (*Tropheops* sp. "mauve"), three benthic (*Aulonocara stuartgranti*, *Otopharynx argyrosoma*, *Copadichromis chrysonotus*), and one pelagic species (two distinct individuals of *Rhamphochromis* sp. "chilingali"). These genomes were sequenced with Pacific Biosciences continuous long-read (PacBio CLR) technology or Oxford Nanopore Technologies (ONT simplex), with details about sample preparation, sequencing, and assembly given in the Methods. In addition, we included previously published genomes for two other species: *A. calliptera*, which lives in rivers and marginal areas around Lake Malawi, and another mbuna, *M. zebra*. Unlike the aforementioned assemblies, these have been scaffolded to chromosome level (scaffold N50 values: *A. calliptera*, 38.7 Mbp; *M. zebra*, 32.7 Mbp) and have gene annotations. Details about the assemblies are summarized in Table 1, and all of them are collapsed, single-haplotype assemblies, meaning that we only have information from one of the paternal or maternal alleles at heterozygous sites.

The contig N50 values of the newly sequenced genomes range from 0.6 to 8.4 Mbp, with the PacBio assemblies showing better contiguity than the ONT counterparts. Although the assemblies are not chromosome level, they score well when assessed for genome integrity by BUSCO v5.5.0 with the OrthoDB "actinopterygii_odb10" data set (Manni et al. 2021). An average of 3580 out of the 3640 essential genes for ray-finned fishes (98.4%) were completely matched in the PacBio assemblies, comparable to 3539 (97.2%) for *A. calliptera* and 3580 (98.4%) for *M. zebra*, respectively. The ONT assemblies for *O. argyrosoma*, *C. chrysonotus*, and *R.* sp. "chilingali" fared worse, with 3406 (93.6%), 3366 (92.5%), and 3420 (93.9%) genes detected, which was expected given that they are more fragmented (Supplemental Fig. S1). Nevertheless, these BUSCO scores suggest that our assemblies exhibit overall good integrity when measured by gene completeness. We also observed little overlap misses between them, with only 30 (0.82%) BUSCOs undetected in more than five assemblies (Supplemental Fig. S1). Together, the eight assemblies represent all but one (*Diplotaxodon*) of the six major ecomorphological groups of Lake Malawi cichlids, allowing a good baseline sampling of the genomic information in the Lake Malawi radiation.

We utilized the minigraph package to construct a pangenome graph from the Lake Malawi haplochromine cichlid assemblies; minigraph successively adds new genome sequences to the graph structure by a genome-to-graph alignment, starting from a reference genome assembly, known as the backbone (Li et al. 2020). Bubbles are augmented onto the existing graph to represent regions where the new assembly differs sufficiently in sequence from all those already incorporated. We chose the *A. calliptera* (astCal) genome as the backbone because it is the most contiguous assembly with Ensembl gene annotations available, allowing us to define where SVs are positioned relative to known genes. The remaining assemblies were incorporated in the following order, prioritizing phylogenetic proximity and genome quality: *M. zebra* (mayZeb), *Tropheops* sp. "mauve" (troMau), *A. stuartgranti* (aulStu), *O. argyrosoma* (otoArg), *C. chrysonotus* (copChr), and the two *Rhamphochromis* sp. "chilingali" (rhaChi, rhaChi2).

Structure of the Lake Malawi cichlid pangenome graph

The Lake Malawi cichlid multiassembly graph consisted of linear structures representing the backbone scaffolds, punctuated by bubbles representing SVs (Fig. 1B). Overall, the graph contains 1.171 Gbp distributed across 637,237 segments connected by 913,087 edges. On average, a segment measured 1,839.04 nucleotides in length and was connected by 1.43 edges. The linear part of the graph comprised 189,197 segments containing 758.85 Mbp (64.8% of the total), whereas the variable part (bubbles) was made of 448,040 segments containing 413.05 Mbp (35.2%). As the query assemblies were incorporated, the graph showed greater complexity with increases in the number of edges and segments, accompanied by a decrease in the average length of segments

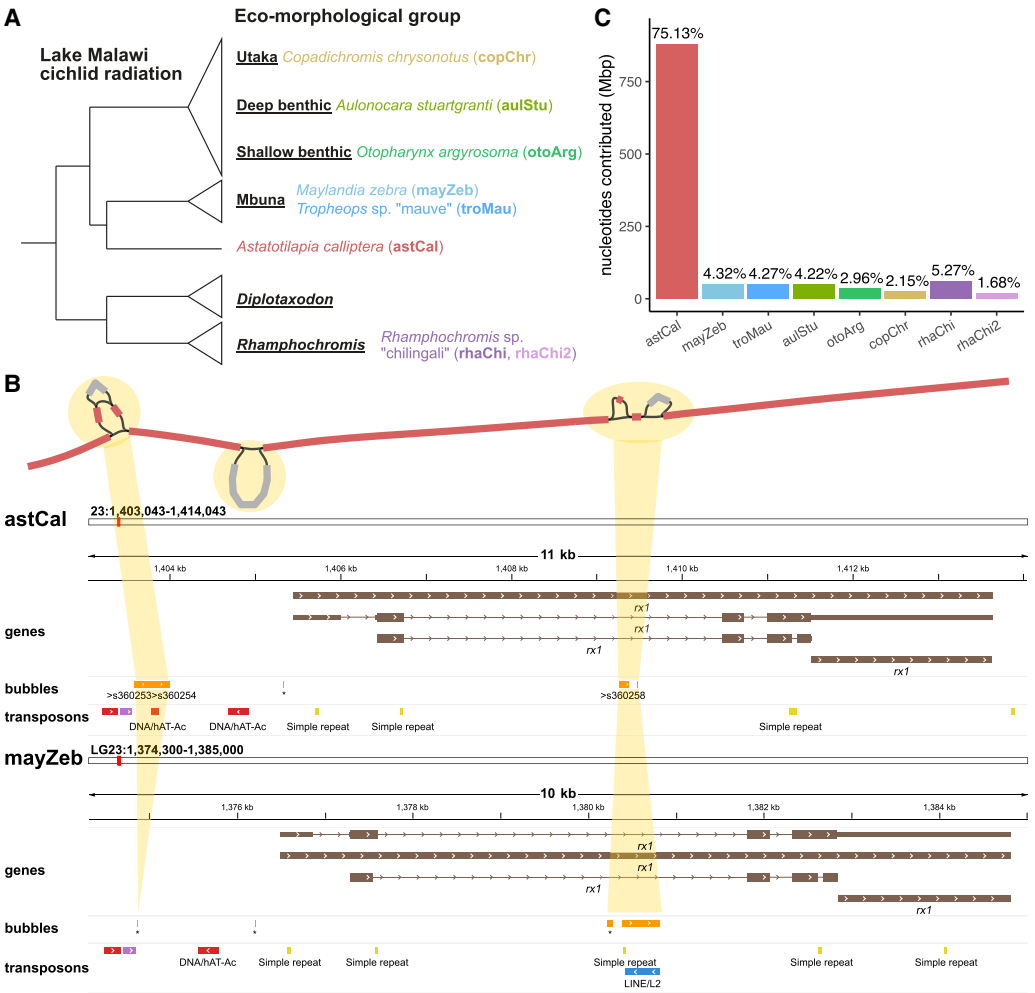

Figure 1. Features of the Lake Malawi haplochromine cichlid assemblies and pangenome graph. (*A*) Phylogenetic tree of cichlid species used in this study, shown as part of their ecomorphological groups. (*B*) Visualization of the pangenome graph around the retinal homeobox 1 (*rx1*) gene involved in cichlid opsin expression. Structural variants (SVs) are represented as bubbles along the *Astatotilapia calliptera* fAstCal1.2 backbone, whose segments are colored red. These bubbles are juxtaposed with their corresponding locations on the *A. callitpera* and *Maylandia zebra* linear reference genomes, on which gene and transposon annotations are provided. (*C*) Percentage of sequence in the pangenome contributed from each assembly.

(Supplemental Fig. S2). The backbone-guided nature of graph construction in minigraph meant that the majority of the graph segments originated from the *A. calliptera* backbone, whose 880,428,986 nucleotides were encompassed within 382,935 segments (Fig. 1C). Incrementally incorporating the other assemblies grew the reference graph by the corresponding number of segments: mayZeb (62,307), troMau (40,627), aulStu (44,206), otoArg (29,209), copChr (21,444), rhaChi (40,742), and rhaChi2 (15,767). These segments consisted of varying number of nucleotides: mayZeb (50,651,908), troMau (50,003,806), aulStu (49,458,672), otoArg (34,661,015), copChr (25,161,891), rhaChi (61,804,019), and rhaChi2 (19,731,824). The query assemblies contributed a cumulative total of 291.47 Mbp, or 33.1%, additional nucleotides compared with the *A. calliptera* backbone. These nonreference sequences were encapsulated within a total of 188,944 bubbles, positioned in backbone regions that span the structural breakpoints from at least one of the nonreference samples. Out of the 880,428,986 nucleotides on the backbone segments, a substantial 840,567,200 (95.47%) received spanning coverage (Supplemental Figs. S3, S4), which translates to a bubble density of 0.2248 per 1 kbp of covered backbone sequence.

The construction of a multiassembly graph with minigraph relies on the initial choice of a backbone assembly, and it has been reported that this could directly impact the amount of nonreference sequence that is detected from the subsequent assemblies (Crysnanto et al. 2021). To investigate the robustness of our multiassembly graph, we built seven other graphs using each of the nonreference samples as the backbone in turn. The more contiguous PacBio backbones produced graphs that were topologically more complex than their ONT counterparts with larger number of segments, edges, bubbles, and extra sequence. Nevertheless, the proportions of linear and variable sequences were similar across all the backbones at ~65% and ~35%, respectively (Supplemental Fig. S5), whereas most metrics in the PacBio-based graphs fell within 5% of those of the canonical *A. calliptera* version (Supplemental Table S1). Permuting the order of incorporation of the subsequent assemblies did not significantly affect graph structure and topology, suggesting that backbone choice was the dominant factor in

Table 1. Lake Malawi haplochromine cichlid genome assemblies

Species/sample	Short name	Sex	Sequencing technology	Assembler	No. of contigs/ scaffolds	Contig/ scaffold N50 (Mbp)	Total size (Gbp)	Sequencing depth	BUSCO score
Astatotilapia calliptera	astCal	?	PacBio CLR + Illumina	Falcon	738 (248)	4.4 (38.7)	0.88	80.72	97.2
Maylandia zebra	mayZeb	Male	PacBio + Illumina	Falcon	2331 (1690)	1.4 (32.7)	0.96	154.12	98.4
Tropheops sp. "mauve"	troMau	Male	PacBio CLR	Falcon	271	8.4	0.91	137.72	98.4
Aulonocara stuartgranti	aulStu	Male	PacBio CLR	Falcon	1027	2.1	0.89	86.31	98.3
Otopharynx argyrosoma	otoArg	Male	ONT	Shasta	2861	2.5	0.88	24.44	93.6
Copadichromis chrysonotus	copChr	Male	ONT	Shasta	6225	0.6	0.86	19.43	92.5
Rhamphcchromis sp. "chilingali"	rhaChi	Male	PacBio CLR	Falcon	437	4.4	0.90	94.37	98.4
	rhaChi2	Female	ONT	Shasta	3239	1.2	0.85	42.13	93.9

astCal and mayZeb are previously published chromosome-level assemblies from Ensembl. Values in parentheses for these two indicate properties calculated for scaffolds, in addition to the contig-level statistics like for the other assemblies.

the variation that was observed (Supplemental Fig. S6). For a fair comparison of bubble frequency, we normalized the bubble counts by the number of backbone nucleotides with spanning coverage, and found that these values ranged from 0.2192 to 0.2254 bubbles per kilobase pair of covered genomic sequence (*A. calliptera*, 0.2248). These observations suggest that minigraph is consistent for SV discovery and that the default graph exhibits a reasonable structure and amount of sequence information.

Finally, we examined how multiassembly graph construction behaved in response to the choice of the minimum variant length (L) in minigraph, which influences the minimum length of sequence difference required for a bubble to be created. We utilized the default value of L = 50, as it follows the conventional definition of a SV. This has the advantage of ignoring smaller-scale variation (SNPs, small indels) and making the graph less complex and more interpretable, which is appropriate for the focus on larger-scale variation in this study. Although the topological complexity of the multiassembly graph decreases with larger values of L, we found that there was a relatively broad parameter range of L = 25 to 500 at which the total sequence content in the reference graph remained relatively stable (Supplemental Fig. S7; Supplemental Table S2).

Substantial proportion of nonreference sequences in the pangenome

The majority of SV sequences in the graph bubbles of the Lake Malawi cichlid pangenome originated from nonreference assemblies rather than the backbone (Fig. 2A). The amount of sequence from the seven other assemblies totaled up to 291.47 Mbp, or 33.1% of the size of the *A. calliptera* backbone, averaging to a median of 49.46 Mbp per assembly. Conversely, the amount of *A. calliptera* backbone sequence located within bubbles was lower at ~13.81%, or 121.57 Mbp. Repeating this analysis across the graphs with different backbones also showed a similar pattern in which more of the uncovered SV sequences were attributed to nonreference sequences rather than the backbone (Supplemental Fig. S8). This suggests that each assembly contains stretches of genomic sequences that might have been overlooked if we had only focused on any individual backbone as a reference, highlighting the value of the pangenome approach.

Following this, we uncovered that the nonreference sequences in the pangenome did not originate from a single assembly, but instead, similar amounts were contributed by the seven genomes. The majority of nonreference graph segments containing the SV sequences were only present in one genome, making them singletons or private variation (Fig. 2B,C). These cumulatively account for 182.16 (62.5%) out of 291.47 Mbp spread across 115,173 (45.3%) out of the 254,201 nonreference segments within the bubbles. Notably, the higher-quality PacBio assemblies contributed the most unique sequences to the pangenome: troMau (33.98 Mbp), aulStu (30.49 Mbp), and rhaChi (28.99 Mbp). However, we also observed some degree of sequence common between the genomes (Fig. 2B,C), the greatest amount of which was between the two *Rhamphochromis* (28.09 Mbp shared across 21,272 segments). Finally, there was 4.98 Mbp of sequence detectable in all the nonreference assemblies but absent from *A. calliptera*. These bubbles instead contained a substantial amount of backbone sequences that were private to the *A. calliptera* assembly: 22,511 segments harboring 42.32 Mbp (4.81% of the backbone). These results suggest that the SV landscape of the Lake Malawi cichlids is dominated by singletons.

Discovery of SVs

We investigated the 188,452 bubbles in the reference graph in detail and found that the majority harbored straightforward presence-or-absence variation or slightly nested variation, with complex structural variation being relatively rare (Supplemental Fig. S9). The number of possible paths to traverse a bubble exhibited a strongly left skewed distribution toward simple variation: There were 151,761 (81%) two-path and 21,230 (11%) three-path bubbles, and 182,136 (97%) had no more than eight theoretical paths (Supplemental Fig. S10). To focus on true biological variation within the bubbles, or the alleles, we retain only the paths actually transversed by samples. Using minigraph, we were able to successfully call alleles for all samples at 160,578 out of the 188,452 bubbles (85.2%). For subsequent analyses, we retained

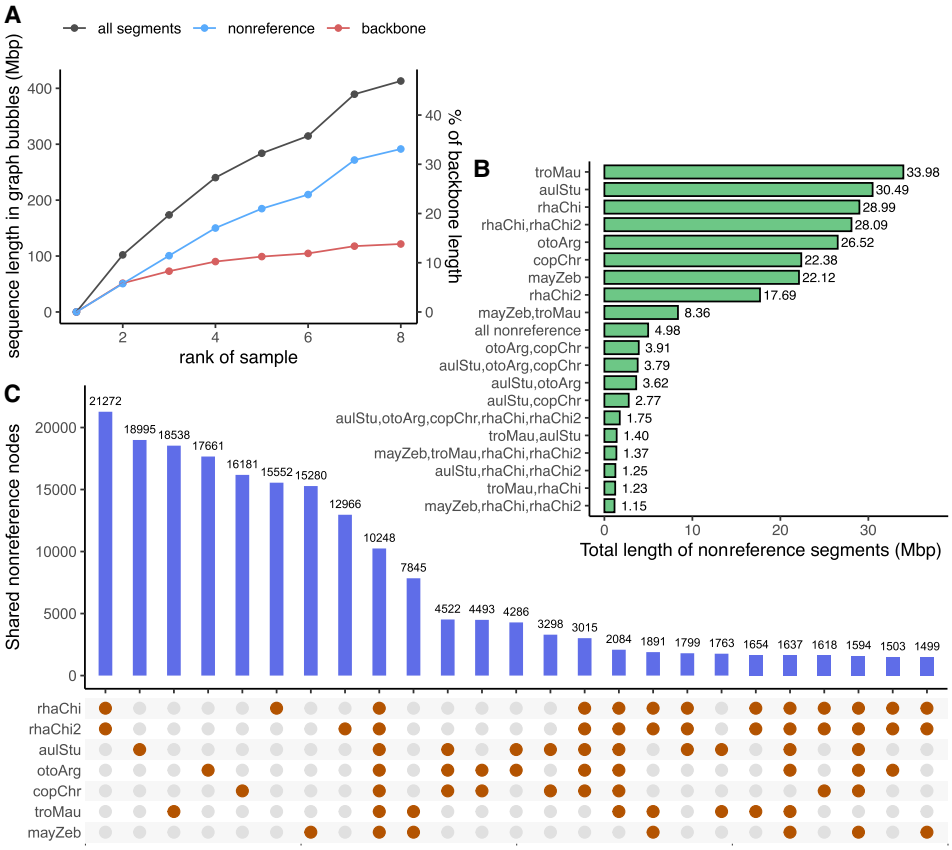

Figure 2. Nonreference sequences in Lake Malawi pangenome graph. (*A*) Length of SV sequences in the graph bubbles, based on whether they originated from nonreference assemblies (blue) or the backbone (red). (*B*,*C*) Cumulative length and number of graph segments shared across assemblies.

187,552 out of the 188,452 bubbles (99.5%) for which the alleles are known for at least five out of eight samples to facilitate between-sample comparisons (Supplemental Fig. S11).

Most of the retained bubbles were biallelic (153,848, 82%), whereas the remaining multiallelic ones mostly contained three alleles (20,751, 11%). The graph was composed of mostly simple structural variation, many of which were singletons (Fig. 3A), but the size of the SVs varied widely. A total of 434,357 alleles were observed across the 187,552 bubbles, which translated to an average of 2.32 alleles per bubble. The 246,805 nonreference alleles in this set comprised 144,769 insertions, 101,179 deletions, and 857 substitutions/inversions with an average length of 2016.7, 1510.8, and 605.2 bases, respectively (Fig. 3B). The cumulative length of insertions was 291.96 Mbp, which was longer than the 152.86 Mbp associated with deletions. The length of SVs shows a peak above 1000 bp, with the top 1% longest insertions ranging from 17,469 to 251,778 bp and deletions measuring 19,567 to 215,374 bp. The Lake Malawi cichlid multiassembly graph contained more complete insertions (78,479; only nonreference allele with sequence, reference length of zero) compared with partial/alternate insertions (66,290; both reference and nonreference sequences present, but the nonreference allele is longer). We observed the opposite pattern with deletions: 49,217 complete versus 51,962 partial/alternate.

Most of the biological alleles in the reference graph are single insertion or deletion events that are present at a low sample frequency across our eight assemblies. A significant majority of alleles (147,802 or 60%) are singletons, meaning that the nonreference allele had a sample frequency of either one out of eight (135,160 alleles), or seven out of eight (12,642 alleles), in which case the reference allele on the *A. calliptera* backbone is the singleton (Fig. 3C). These results are consistent with the observation in the previous section that many nonreference segments and sequences are private to a single assembly (Fig. 2). The extreme sample frequencies were primarily dominated by the large SVs measuring thousands of base pairs in length, with one of eight and two of eight frequencies dominated by complete insertions and seven of eight by complete deletions (i.e., complete insertions in *A. calliptera* backbone relative to the others). Conversely, intermediate allele frequencies tended to be composed of shorter sequence variants, with most having a length of ~50 bp (Fig. 3C).

Recapitulation of species relationships

We investigated how the phylogenetic information in the multiassembly graph compares to what is known about species relationships from SNP studies of the Lake Malawi cichlid radiation (Malinsky et al. 2018). Overall, the pangenome graph closely reflected the expected interspecies relationships based on the prevailing understanding of the evolutionary history of the Lake Malawi cichlid radiation (Fig. 1A), despite the fact that the assemblies varied in their quality and the sequencing technologies used.

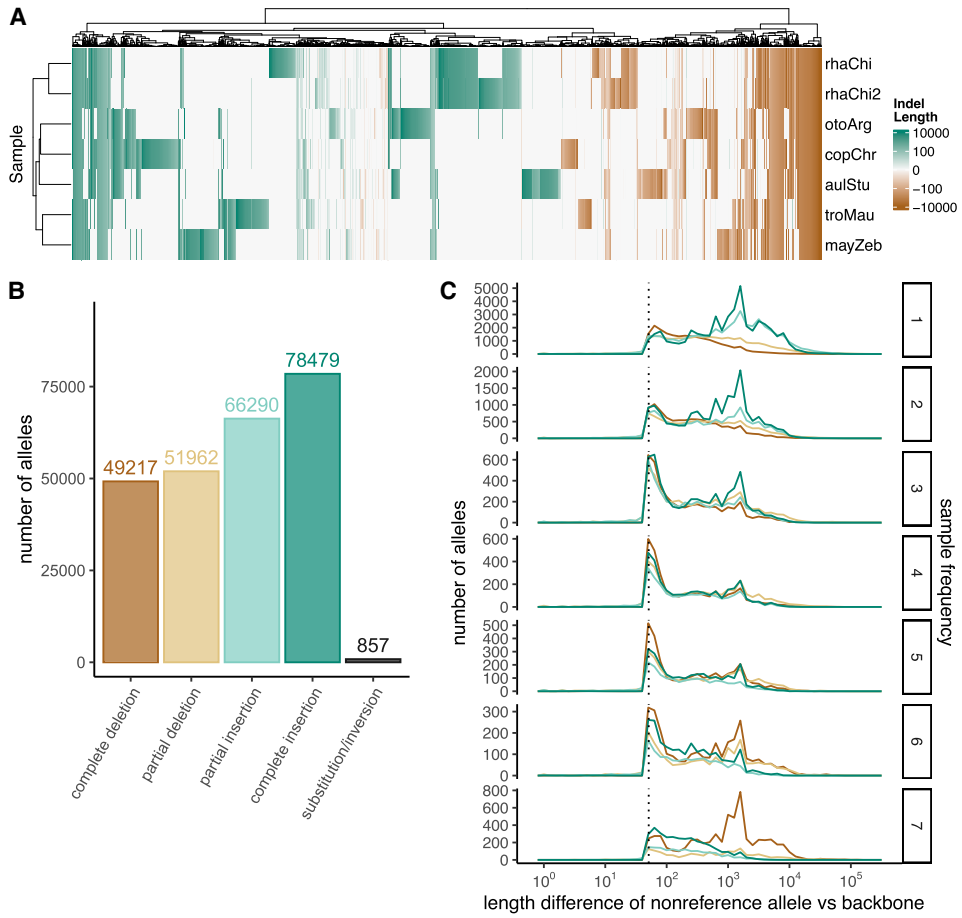

Figure 3. The Lake Malawi cichlid SV landscape. (*A*) Heatmap showing length difference of alleles for each nonreference sample versus the *A. calliptera* backbone across 5000 random bubbles, shown as columns. A positive value (green) represents an insertion, and a negative value (brown) represents a deletion, with darker colors indicating larger SVs/indels. (*B*) Number of SV alleles by type. (*C*) Length deviations of nonreference alleles across different sample frequencies. Dashed line denotes 50 bp.

For instance, we successfully recapitulated known cichlid groupings in a principal component analysis (PCA) plot based on a presence/absence matrix across the 150,532 bubbles containing two or three observed alleles (Fig. 4A). Using this same presence/absence matrix, we also constructed a phylogenetic tree based on the number of SV events between samples (Fig. 4B), revealing a median of 56,402 differences. We also observed a considerable amount of interspecies polymorphism, with the highest levels between the mbuna species *Tropheops* sp. "mauve" and *M. zebra*, which shared 111,695 events (74.2%). The two *Rhamphochromis* individuals shared 124,364 (82.6%), while also showing higher correlation of SV lengths and differing allele sequences at noticeably fewer positions (Supplemental Fig. S12).

Next, we constructed biassembly graphs from every possible pair of samples to perform a one-to-one estimation of the sequence that manifests as SVs between them. This was intended to complement previous estimates of 0.1% to 0.25% pairwise SNP divergence in Lake Malawi cichlids. Unlike SNP studies, this comparison is asymmetrical, in which one assembly acts as a backbone and the other as a query, from which the SV percentage is estimated as the percentage of backbone regions within bubbles (Fig. 4C). With the default minimum variant length parameter L = 50 and correcting for sequencing coverage, estimated interspecies SV proportion ranged from 4.73% to 9.86% (mean, 7.11%), with smaller values within the mbuna (mean, 5.45%) and benthic (mean, 5.20%) groups. The estimated differences were much lower within species when comparing the two *Rhamphochromis* individuals (2.96% and 3.68%). We tested parameter values of L from 2 to 250 and found that these estimated percentage differences were representative of those at larger L values, as smaller variants and potential sequencing errors in the assemblies were gradually excluded (Supplemental Fig. S13).

Using these same biassembly graphs, we next obtained a measure of how frequently structural events occur along the genome by calculating the bubble density normalized by spanning coverage (Fig. 4D). This revealed an interspecies bubble density of 0.061 to 0.098 per 1 kbp of backbone compared with an intraspecies density of 0.041 among the two *Rhamphochromis*. There appeared to be some directionality to the sequence changes, as both *Rhamphochromis* individuals contained more insertions and fewer deletions when aligned to the other species (Supplemental Fig. S14). This high similarity between the two *Rhamphochromis* individuals is expected as they belong to the same species and might have exhibited structural changes independently of the other

cichlids. This is consistent with the SNP-based Lake Malawi cichlid phylogeny in which the pelagic group that contains the *Rhamphochromis* and *Diplotaxodon* taxa branched off earlier than the rest of the radiation (Fig. 1A).

PCR validation and genotyping

We performed polymerase chain reaction (PCR) validation of 16 predicted SVs, using five of the original tissue samples that were used for genome sequencing of troMau, otoArg, copChr, rhaChi, and rhaChi2. Because of difficulties in designing PCR primers for repetitive genomic regions, we prioritized bubbles containing simple, straightforward presence-or-absence variants located within 1500 bp of a gene. Out of the bubbles we tested, the validation results for 12 were consistent with the allele predictions for all five assemblies, whereas the remaining four matched the allele predictions of four assemblies (Fig. 5A; Supplemental Table S3). Some heterozygosity was present in four of the 12 bubbles, which was expected because our samples were not from homozygous lines. At such heterozygous sites, allelic information across parental chromosomes would have been collapsed into a single sequence, showing only one of two alleles in the assembly. Ten out of 16 SV PCR results were further verified by Sanger sequencing (Supplemental Table S3). For the remaining six out of 16 SVs, PCR results were partially validated by Sanger sequencing, as we could not acquire good quality sequencing in some instances, likely because of the repetitive nature of these loci. However, all sequenced PCR products that were obtained corresponded to the expected sequences. Overall, these results provide additional experimental support to our findings and demonstrate the reliability of SV detection with minigraph.

We also utilized the designed PCR primers to perform additional genotyping of a wider cohort of aquarium-reared male and female individuals of *A. calliptera*, *Tropheops* sp. "mauve" and *Rhamphochromis* sp. "chilingali" to check for evidence of polymorphism at some bubbles (Fig. 5B). For example, we did not find any evidence for polymorphism in a bubble near *mitfa* (ENSACLG00000022427), with all genotyped *A. calliptera* and *Rhamphochromis* sp. "chilingali" showing one allele and all *Tropheops* sp. "mauve" showing the other. However, bubbles near some genes (*mfsd4a* and ENSACLG00000002767) showed evidence of polymorphism in different sexes and species. These observations suggest a variable degree of SV polymorphism and heterozygosity in Lake Malawi cichlids. However, there is limited statistical power to draw any conclusions about the wider Lake Malawi cichlid radiation given our small sample sizes.

SVs are underrepresented in protein-coding regions

To investigate the functional relevance of the SVs, we examined the 187,552 bubbles for the presence of nearby genes. We focused on a subset of 26,734 out of 28,001 gene annotations in the *A. calliptera* genome, filtering out putative TEs misannotated as genes (see Methods); 90,531 (48.3%) bubbles directly intersected with genic features like exons, introns, or UTRs, whereas 8967 (4.8%) and 6010 (3.2%) were located within 2000 bp upstream of and downstream from genes, respectively. The remaining 82,024 (43.7%) were considered intergenic, located an average of 38.4 kbp from the nearest gene start. Coverage-corrected bubble

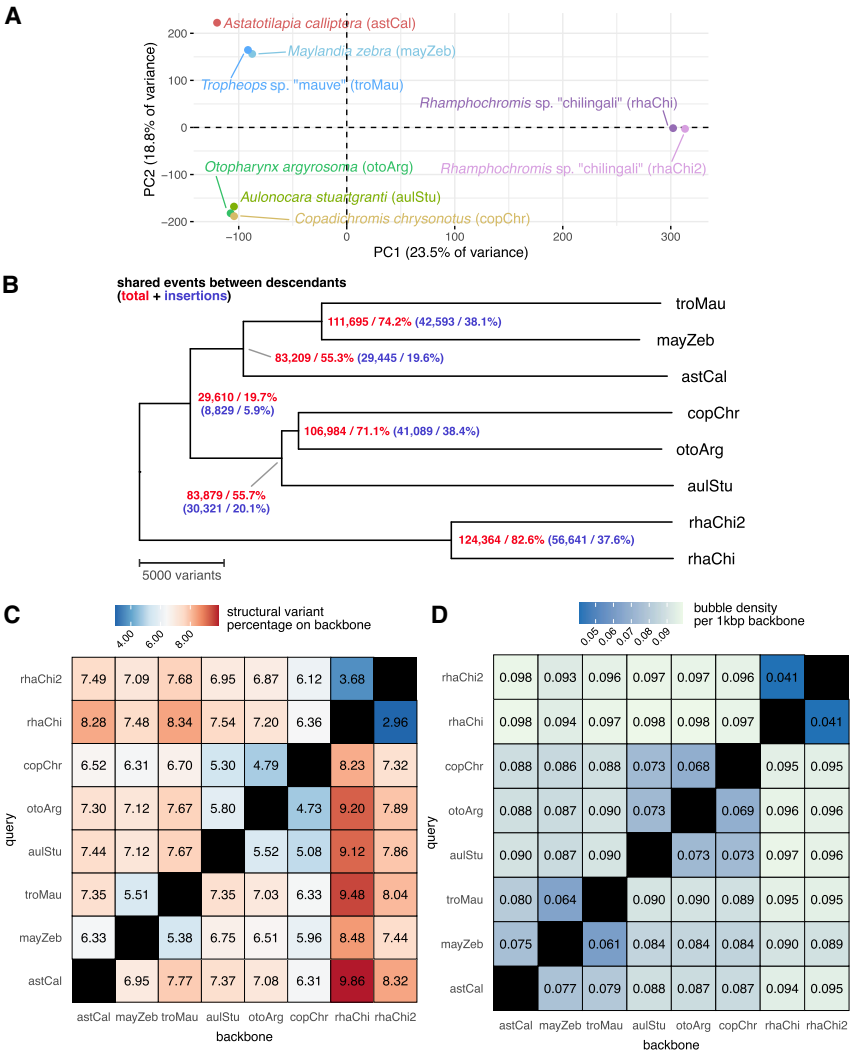

Figure 4. Estimation of phylogenetic relationships between Lake Malawi cichlid assemblies. (*A*) Principal component analysis of the samples based on a presence/absence matrix across 150,532 bubbles with at most three alleles. (*B*) Midpoint-rooted neighbor joining tree of Malawi samples estimated from SV events. The distance between any two samples corresponds to the total horizontal branch length separating them. Nodes are labeled with the number of sites at which descendant samples share the same allele (first number) and inserted sequences relative to the shortest allele observed across all samples (second number). (*C,D*) Structural variation manifested between biassembly graphs between sample pairs as a percentage or bubble density. Comparisons are asymmetrical, in which one assembly acts as a backbone against which the other query is aligned. Minimum variant size in minigraph was set to 50.

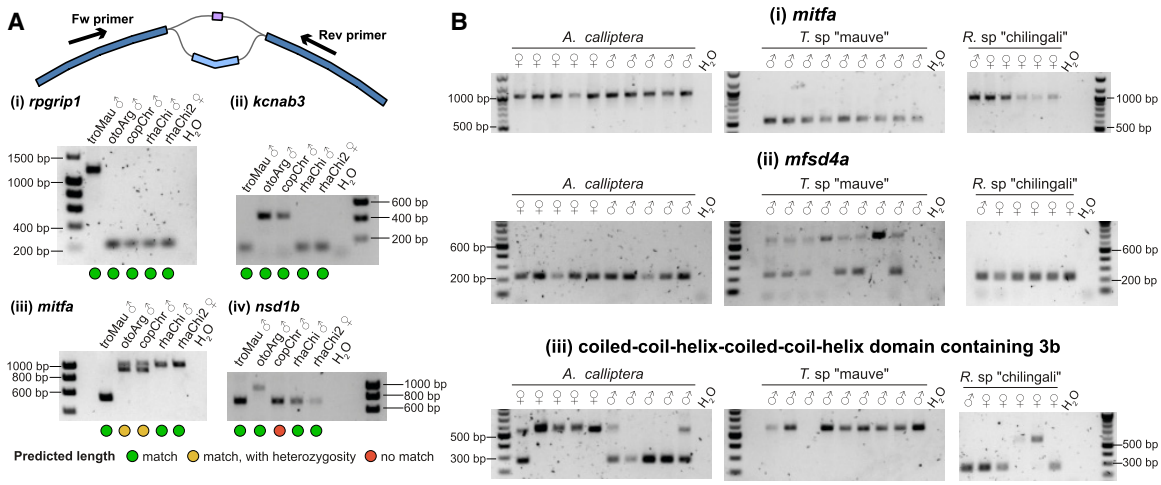

Figure 5. PCR genotyping of SVs. (*A*) Validation of predicted SVs in original samples. Green indicates match; amber, match with heterozygosity; and red, no match. (*B*) Wider genotyping across aquarium-grown individuals: 10 *A. calliptera* (five males, five females), nine *Tropheops* males, and six *Rhamphochromis* (one male, five females).

densities were 0.2481 per kilobase pair of intergenic sequence and 0.2213 for introns, similar to the genome-wide level of 0.2248, but the bubble density in exonic regions was significantly lower at 0.0899 per kilobase pair (Fig. 6A). Functional enrichment analysis found that genes containing SVs are associated with broad Gene Ontology categories, including transport (P-value = 2.122×10^{-33}), phosphorus metabolic process (P-value = 6.204×10^{-23}), and developmental process (P-value = 5.660×10^{-21}) (Supplemental Fig. S15). The transport and phosphorus metabolic process categories include potassium channels and bone morphogenetic proteins, whereas the developmental process involves forkhead box (FOX) genes and fibroblast growth factor receptor 1. There were 4723 (17.7%) genes in which SVs were entirely absent from the gene body and within a 2000 bp region, which we consider a highly conserved set of genes across species. This gene set was associated with terms related to transcription and translation: gene expression (P-value = 4.4×10^{-40}), RNA-induced silencing complex (RISC) (P-value = 9.2×10^{-38}), DNA-binding transcription factor activity (P-value = 1.6×10^{-13}), and structural component of ribosome (P-value = 1.9×10^{-10}) (Supplemental Fig. S16).

Subsequently, we estimated the percentage conservation of genes for each sample based on the path they traversed through the graph. This calculation was performed for a set of 20,785 (86.9%) genes on the *A. calliptera*–backboned graph, which were located outside overly complex bubbles and had spanning coverage from all samples. This approach suggested that 83.2% to 85.6% of genes showed >95% sequence conservation in the gene body across the nonbackbone samples, with ecologically similar species exhibiting similar percentage conservation scores to each other (Fig. 6B). This conservation was more pronounced in exonic regions, where 93.5% to 94.7% of genes maintained this high level of conservation. We observed similar patterns with the corresponding set of 20,266 (85.6%) *M. zebra* genes, suggesting this high sequence conservation was robust against reference bias (Supplemental Table S4).

We counted the presence of each gene across the eight assemblies, applying a lenient 50% conservation threshold to accommodate large introns. This analysis covered a larger set of 24,322 (86.9%) *A. calliptera* and 24,532 (85.6%) *M. zebra* genes within regions that had spanning coverage from at least six samples. As expected, a large proportion of genes—21,407 (88.0%) *A. calliptera* and 20,964 (85.5%) *M. zebra*—was detectable in the other assemblies. This approach indicated that *A. calliptera* and *M. zebra* had 378 and 260 gene sequences that were unique relative to the other assemblies (Fig. 6C). These genes contain protein domains that are ubiquitous and multicopy across eukaryotic protein families (Supplemental Fig. S17), including ubiquitin-like, zinc-fingers, ankyrin repeats, and G protein–coupled receptor (GPCR) domains. However, a TBLASTN search suggested that these were likely false positives: 365 (96.6%) supposedly unique *A. calliptera* genes were detectable in *M. zebra*, and 219 (84.2%) vice versa. For instance, several hemoglobin genes thought to be private to the *A. calliptera* were traced to assembly errors that caused the formation of artifact bubbles in the pangenome graph (Supplemental Fig. S18). Overall, the available evidence suggests that the sequence differences between Lake Malawi cichlids are unlikely to be caused by preferential gene gain or loss in certain groups.

Enrichment of TEs in SVs

Having observed that most of the SVs do not occur in protein-coding regions of genes, we turned our attention to TEs, which have been shown to be sources of phenotypic novelty in cichlid fishes (Santos et al. 2014; Carleton et al. 2020; Munby et al. 2021). About 36.55% of the *A. calliptera* sequences are annotated by RepeatModeler/RepeatMasker as TEs, of which the biggest classes were DNA transposons (12.09%), LINEs (8.32%), and LTR retrotransposons (7.38%), whereas the SINEs, helitrons, and retroposons collectively take up <1% (Fig. 7A). There is also a substantial number of unknown elements (8.54%), which might constitute currently uncharacterized TEs and other repetitive elements. Overlap analysis suggested that 74.65% of SVs sequences on the backbone were annotated as TEs, which is a 2.04-fold increase compared with genome-wide TE distribution, or 2.20-fold if unknown elements were excluded (62.25% TEs). This enrichment was still detectable even when SV coordinates were randomized and were across varying sequence divergence thresholds for TE detection in RepeatMasker, with the DNA, LINE, and LTR transposons consistently

showing strong enrichment (Supplemental Fig. S19; Supplemental Table S5). These genome-wide and SV transposon proportions were also reflected in the other species, ranging from 36.32% to 38.29% genome-wide and 72.26% to 75.74% in SVs, suggesting that TEs represent most of the large-scale sequence differences among the cichlid assemblies (Supplemental Table S6).

Closer inspection showed that the multiassembly graph successfully incorporated known examples of SV/TEs in Lake Malawi cichlids, such as a haplochromine-specific SINE insertion upstream of *fhl2b* important for egg spot patterning (Supplemental Fig. S20; Santos et al. 2014), as well as a nested insertion upstream of *rx1* whose alleles contribute to different opsin palettes in cichlid vision (Supplemental Fig. S21; Schulte et al. 2014). Another example is the *mitfa* gene, whose first intron harbors the presence of complex variation caused by TE insertions (Fig. 7B), consistent with previous observations (Carleton et al. 2020). Although we refrain from making claims about species-level differences in TE composition, owing to the varying quality of the assemblies, we performed a high-level characterization of the subclasses of TEs within graph bubbles as they might indicate classes that are still segregating within the Lake Malawi population. By overlapping TE coordinates against SVs across all the assemblies, we identified 254,015 straightforward presence-or-absence TE insertion events across 90,451 unique bubbles (Fig. 7C). The most frequent within-bubble TE insertions involved the subclasses of LTR/Gypsy (38,216), LTR/Unknown (28,080), DNA/TcMar (21,518), DNA/hAT (18,947) LINE/Rex-Babar (18,292), and LINE/L2 (13,077). On a genome-wide scale, 14,868 genes (55.9%) had a putative polymorphic TE bubble in the gene body or within 2000 bp upstream. This number decreased to 7898 (29.7%) if we required the bubble to be in proximity to the start of the gene. None of these intersections revealed a significant Gene Ontology enrichment, even when stratifying for TE classes and subclasses. This lack of functional enrichment supports the hypothesis that species-variable TE insertions display no apparent pattern or preference for specific genes.

Discussion

By integrating eight long-read genome assemblies of Lake Malawi haplochromine cichlids into a multiassembly graph, we uncover and characterize novel structural variation not represented in the established chromosome-level reference assemblies. Although SVs have been studied before in cichlids (Brawand et al. 2014; Fan and Meyer 2014; Conte et al. 2019; Kratochwil et al. 2019; Penso-Dolfin et al. 2020), as far as we are aware, our work represents the first efforts to quantify these large-scale genomic differences in radiating cichlids of an East African lake using a pangenome graph. We estimate that there is 26.4% to 33.1% of additional sequence relative to a chosen assembly (Fig. 2), which is a relatively large amount compared with reports from within-species reference graphs in other vertebrates, including cattle (2.8% in six genomes) (Crysnanto et al. 2021), sheep (~5% in 15 breeds) (Li et al. 2023), duck (2.33% flexible genes in 131 genomes) (Wang et al. 2024), and humans (~10% in 910 African individuals) (Sherman et al. 2019). Direct comparisons of these values are not straightforward owing to variations in methodologies and lack of universal definitions of pangenomic terms (Sherman and Salzberg 2020), and it remains to be seen whether large genomic differences must necessarily correlate with morphological diversity (Plessy et al. 2024). Nevertheless, this study supports the existence of a significant amount of underappreciated sequence

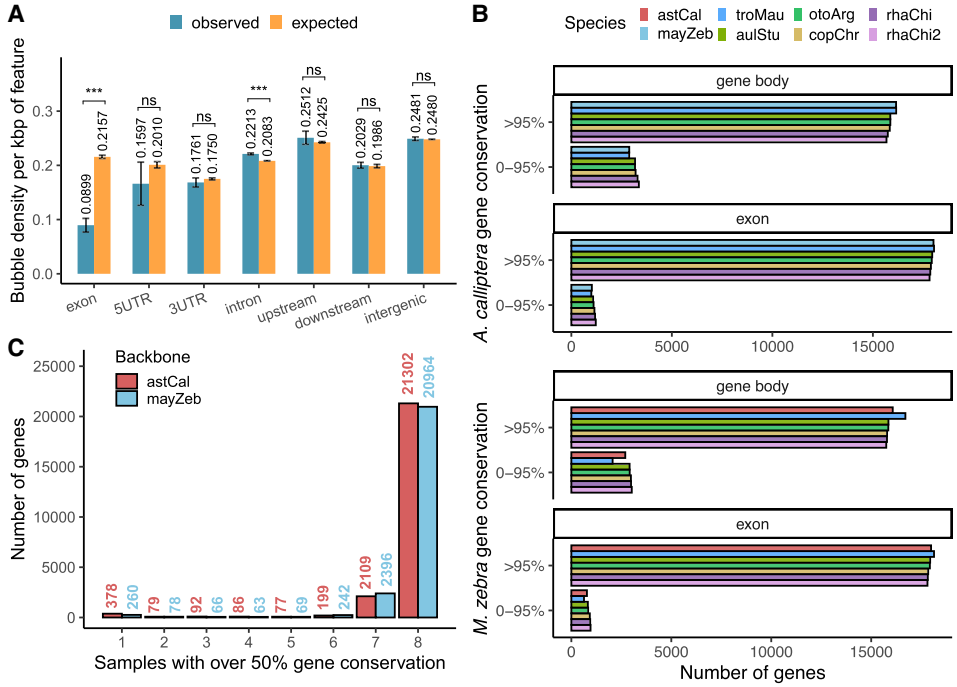

Figure 6. Genomic context of SVs. (*A*) Bubble density in gene features; 95% confidence intervals are estimated across 100 bootstraps, for which the SV coordinates were shuffled within ±1000 bp of their original positions (observed) or randomly across the whole genome (expected). (***) *P*-value < 1×10^{-3}, (n.s.) not significant. (*B*) Percentage sequence conservation of genes calculated for backbones with gene annotations. (*C*) Presence–absence counts of backbone genes. A value of one denotes genes that are not detectable in the other assemblies and might be private, and a value of eight refers to ubiquitous genes.

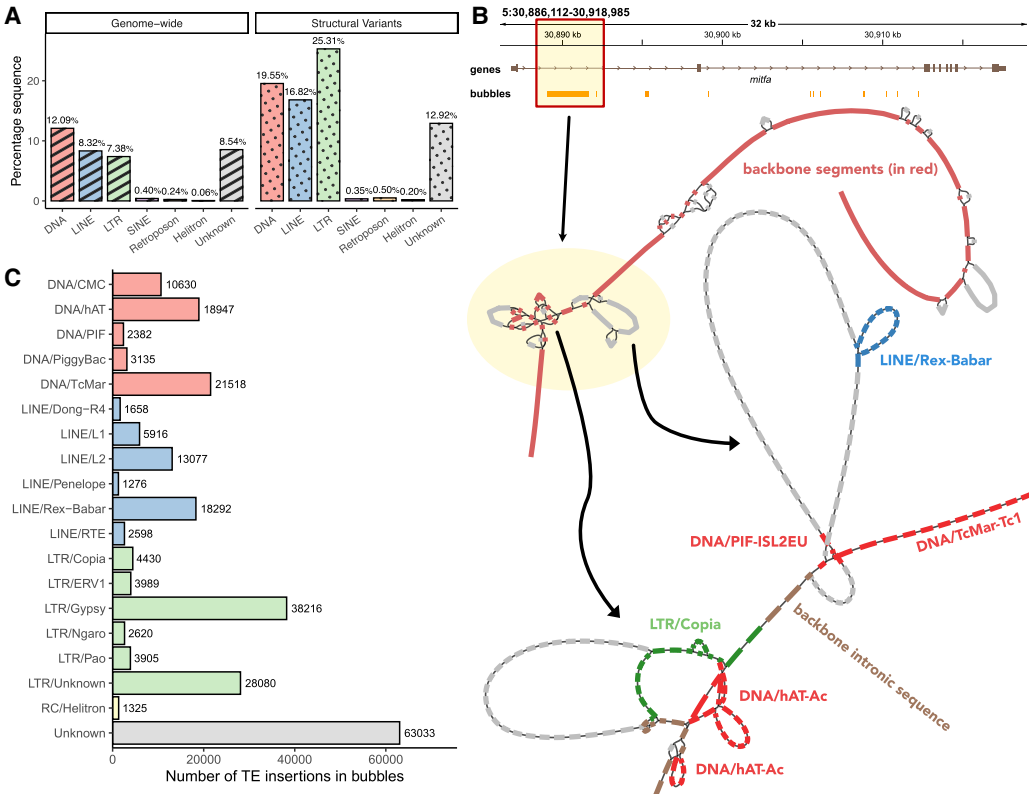

Figure 7. Transposable elements (TEs) in the Lake Malawi cichlid pangenome. (*A*) Genome-wide and SV TE composition of the *A. calliptera* fAstCal1.2 assembly. (*B*) TE insertions in the first intron of *mitfa*, highlighting the presence of two complex bubbles. (*C*) Number of putative polymorphic TE insertion events by subclass across 90,451 bubbles.

diversity in cichlid fishes, and as newer assemblies with improved accuracy become available (Rhie et al. 2021), we anticipate future efforts will characterize SVs in the broader East African cichlid radiation, identifying possible convergent evolutionary patterns across different lake systems.

By taking a genome-wide approach in characterizing SVs, this study complements previous SNP-based estimates of sequence differences between East African cichlid species. Population-scale SNP data had been previously utilized to assess their genetic relatedness (Svardal et al. 2020), with 0.1% to 0.25% pairwise divergences quoted for Lake Malawi cichlids (Malinsky et al. 2018). In contrast, this study primarily focuses on phylogenetic comparisons of long-read assemblies, estimating that 4.73% to 9.86% of base pairs manifest as interspecies cichlid structural variation (Fig. 4). It is worth bearing in mind that SVs and SNPs are not directly comparable as they unveil distinct aspects of genome variation. However, even with a sparse sampling of eight individuals across seven species, these findings indicate the presence of large stretches of hitherto-uncharacterized sequences, which could themselves harbor SNP variation. Although many of the SVs appear as singletons, there is evidence of polymorphism within and between species based on PCR genotyping (Fig. 5), and they contain phylogenetic signals that can separate known ecomorphological groups (Fig. 4C). However, it remains unclear whether the SVs follow a similar population structure and genetic diversity as SNPs, as has been observed for Atlantic salmon (Lecomte et al. 2024). An alternative scenario is that they follow different evolutionary trajectories like in European starling, in which SNPs and SVs are subject to different levels of balancing selection (Stuart et al. 2023), or in capelin fish, in which thermal adaptation is facilitated by copy number variants showing a different demographic history to SNPs (Cayuela et al. 2021). Future cichlid studies should obtain population-level SV information to better assess SV allele frequency distributions. This could potentially lead to the identification of causative variants linked to phenotypic traits through genome-wide association studies or selection scans, as have been attempted with SNPs (Malinsky et al. 2015; Kratochwil et al. 2022).

In light of the growing recognition of the role TEs play in cichlid adaptation and speciation (Santos et al. 2014; Carleton et al. 2020; Munby et al. 2021), this study demonstrates the potential of the pangenome graph as a computationally efficient method to compare multiple long-read assemblies and characterize polymorphic TE insertions at an unbiased, genome-wide level (Ebler et al. 2022; Groza et al. 2024; Igolkina et al. 2024). Although prior research has uncovered isolated instances of potentially beneficial TE insertions in cichlids, the widespread occurrence of TE insertions in the Lake Malawi cichlid genomes, accounting for as high as 74.65% of SV sequences, might provide the evolutionary substrate to produce diverse phenotypes and contribute to cichlid ecological versatility (Ngoepe et al. 2023) through the selection of beneficial polymorphisms (Cayuela et al. 2021; Fang and Edwards 2024) or might simply result in genetic incompatibilities owing to TE accumulation (Mérot et al. 2023). The underrepresentation of these TE insertions from coding regions of genes is not surprising, as such alterations can be highly detrimental by disrupting the reading frame or introducing

premature stop codons (Almeida et al. 2022). It is intriguing to speculate whether persistent TE activity might serve as a mechanism of enhancing the potential for functional genomic differences among closely related species, even in the absence of SNP differences. In conclusion, our findings underscore the complex interplay of evolutionary forces shaping the Malawi cichlid genome, which is likely the product of not only SNPs and small-scale differences caused by point mutation but also of structural variation derived from TEs.

Methods

Genome assemblies

A total of eight Lake Malawi cichlid long-read genomes were included: two previously published chromosome-level assemblies from Ensembl v103, *A. calliptera* (fAstCal1.2, GCF_900246225.1) (Rhie et al. 2021) and *M. zebra* (M_zebra_UMD2a, GCA_000238955.4) (Conte et al. 2019), as well as six contig-level assemblies generated for five other species from aquarium-grown fishes —PacBio CLR for *Tropheops* sp. "mauve," *A. stuartgranti* (male), and *Rhamphochromis* sp. "chilingali" (male) and ONT simplex for *O. argyrosoma* (male), *C. chrysonotus* (male), and *Rhamphochromis* sp. "chilingali" (female). Full details about the wet lab experimental protocol and computational methods for genome assembly are available in the Supplemental Methods. Evaluation of genome properties (N50, contig count, etc.) was performed with QUAST v5.2.0 (Gurevich et al. 2013). Sequencing depth of the read sets was approximated by counting the total number of sequenced bases for each read set using the program composition (https://github.com/richarddurbin/rotate) and dividing that number by the total nucleotide length of the most contiguous genome assembly (*A. calliptera*, 880,445,564 bp). Genome completeness was evaluated with BUSCO v5.5.0 (Manni et al. 2021) using the "actinopterygii_odb10" data set from OrthoDB (Manni et al. 2021).

Graph construction

The cichlid assemblies were integrated into a multiassembly graph using minigraph v0.18-r538 with the graph generation `-xggs` preset. Base alignment `-c` was activated, and the minimum variant length `L` set to the default 50. For the canonical graph, the *A. calliptera* fAstCal1.2 assembly was utilized as the backbone, on which the remaining species were incorporated to create bubbles of structural variation. Separate graphs were also generated using the seven other Malawi assemblies as backbones for comparisons. For each choice of backbone, we generated an additional 30 different random permutations of incorporating species to examine variability in the graph structure and topology. For the canonical version, we also tested values of 1, 2, 5, 10, 25, 100, 250, 500, and 1000 for the minimum variant length `L`. Empirical properties of genome graphs (total nucleotide length, segment and edge count, etc.) were obtained by parsing the GFA output files with custom Python scripts and the `gfatools stat` command in gfatools v0.4-r214 (https://github.com/lh3/gfatools). The `gfatools bubble` algorithm was also used to extract bubble coordinates and properties, which were converted into a BED file for linear visualization in the Integrative Genomics Viewer (IGV) 2.16.1 (Robinson et al. 2011). Genome graphs were also visualized in 2D with Bandage v0.8.1 (Wick et al. 2015).

To determine sequencing coverage over graph segments, we aligned individual nonbackbone assemblies onto the graph using minigraph options `-xasm --cov -c` (assembly-to-sequence mapping, coverage, base alignment). The path an assembly traverses through a bubble can only be determined if the sample has "spanning coverage" on the corresponding backbone region in the linear reference bridging from one end of the bubble to the other. The total percentage of backbone nucleotides with spanning coverage by at least one nonbackbone assembly was used as an overall cumulative coverage value in order to normalize certain statistics (e.g., bubble density) for a fairer comparison between different graphs.

In addition, we generated biassembly graphs for every possible pair of the Lake Malawi cichlid assemblies ($8 \times 7 = 56$), allowing a one-to-one comparison without any potential alignment bias caused by the augmentation of the other assemblies. There were two possible graphs for any given pair, because one assembly can act as a backbone to which another (query) was aligned, making this an asymmetrical comparison. The graph construction `-xgen` and coverage calculation `-xasm --cov` were performed as above. The estimated SV proportion was calculated as how much backbone genomic sequence was located within bubbles, divided by the total backbone length with spanning coverage. Other statistics were calculated by parsing the GFA file or using gfatools v0.4-r214.

SV calling and computing sample relationships

The canonical *A. calliptera* backboned graph was utilized for SV calling using minigraph with the option `--xasm --call -c`. This allowed the identification of each sample's allele, or biological path, through every bubble, defined as the sequence formed by the segments in the taken path. We classified SV alleles into four main types, following the method of Crysnanto et al. (2021):

- Complete deletion—only reference sequence present; nonreference has length zero.
- Partial/alternate deletion—reference and nonreference allele present; but nonreference is shorter.
- Partial/alternate insertion—reference and nonreference allele present; but the reference is shorter.
- Complete insertion—only nonreference sequence present; reference has length zero.

The theoretical complexity of a bubble was determined by `gfatools bubble` as the number of possible paths through the nested segments. Biological paths are denoted as those paths traversed by the assemblies, as determined by minigraph SV calling. Bubbles for which the ratio of the shortest divided by the longest biological allele was close to zero were interpreted as straightforward presence-or-absence variation in certain analyses. The presence/absence matrix and allele lengths were used for computing intra- and interspecies relationships using R packages, including phylogenetic trees (APE v5.4) (Paradis et al. 2004), heatmaps (ComplexHeatmap v2.14.0) (Gu 2022), PCA (factoextra v1.0.7) (https://CRAN.R-project.org/package=factoextra), Pearson's correlation (ggcorrplot v0.1.4) (https://CRAN.R-project.org/package=ggcorrplot), and the tidyverse (v2.0.0) suite of packages (Wickham et al. 2019). Analyses were performed using R Statistical Software v4.2.1 (R Core Team 2021).

PCR validation and genotyping

Forward and reverse primers were designed for selected bubbles, such that the PCR reaction extended inward to the bubble, producing different sized products depending on the allele. PCR products were also extracted for Sanger sequencing to confirm their identity. For more details, see the Supplemental Methods.

Genes

The coordinates of structural breakpoints at bubbles were determined according to the *A. calliptera* fAstCal1.2 reference

coordinates, which were superimposed against those of gene annotations. For preprocessing, we filtered out certain gene entries as misannotations, removing entries where either (1) 70% of gene sequence overlap with a single TE fragment, (2) >90% overlap with multiple TE fragments, or (3) annotated with GO term "transposition." We employed two approaches for SV-to-gene association, the first of which involves computing the overlap or proximity of each individual SV to its closest gene and gene feature. Bubbles that coincide with multiple features were counted only once in decreasing hierarchy: coding exon, 5′ UTR, 3′ UTR, intron, upstream (2 kbp), downstream (2 kbp), and intergenic. For the second, gene-centric approach, we examined each gene by counting the number of SVs in its gene body and flanking 2000 bp regions. Both approaches made use of the BEDTools suite (v2.31.0) (Quinlan and Hall 2010) and the GenomicRanges (v1.50.2)+GenomicFeatures (v1.50.4) packages (Lawrence et al. 2013). Gene Ontology functional enrichment analysis for gene sets was performed with the gprofiler2 R package v0.2.2 (Kolberg et al. 2020).

We utilized ODGI v0.7.3 (Guarracino et al. 2022) to estimate the percentage sequence conservation of genes. The minigraph GFA output graph was converted into ODGI's proprietary format .og with the `build`, `sort`, and `chop` functions, after which `odgi pav` was utilized to superimpose gene annotations onto the graph and compute the percentage sequence conservation of each gene as the number of gene segments traversed by a given sample. We performed this analysis for *A. calliptera* and *M. zebra* backboned graphs with the respective Ensembl gene annotations. Gene sequences identified as private to the backbones were validated with TBLASTN (BLAST package v2.14.0, with parameters `-outfmt 7 -evalue 1e-3`) (Camacho et al. 2009) of the *A. calliptera* private proteins to the *M. zebra* genome and, reciprocally, of the *M. zebra* proteins to the *A. calliptera* genome. The protein sequences of these private genes were investigated further by running hmmsearch against the Pfam database Pfam-A.hmm (version 3.1b2, February 2015) (Mistry et al. 2021).

Transposable elements

The annotation of TEs was performed with RepeatModeler v2.0.3 and RepeatMasker v4.1.2-p1, which were included as part of the Dfam TEtools Docker/Singularity container v1.3 (Storer et al. 2021; docker://dfam/tetools:1.3). We generated de novo repeat families using the `BuildDatabase` and `RepeatModeler` commands, with the `-LTRStruct` option activated. These de novo libraries were then combined with Repbase-derived RepeatMasker libraries (Bao et al. 2015), which were then used to annotate all the genomes with RepeatMasker under options `-e rmblast -no_is`. We overlapped coordinates of TEs with graph bubbles using GenomicRanges (v1.50.2) and GenomicFeatures (v1.50.4). To generate a null model to estimate statistical enrichment of TEs in SVs, we generated 100 random shufflings of SV coordinates using BEDTools `shuffle` (v2.31.0) (Quinlan and Hall 2010), with the `-incl` parameter to force the coordinates to maintain proximity to the original and be in regions with sequencing coverage.

Data access

The raw sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJEB80761, PRJEB80765, PRJEB80840, PRJNA1144831, PRJNA1144838, and PRJNA1144843. The new genome assemblies and additional data like the pangenome graph and list of discovered variants are also deposited on Zenodo (https://doi.org/10.5281/zenodo.14029308). The code used for analysis is provided as Supplemental Code. Under an access and benefit sharing agreement, these data are made available on an open access basis for research use only. Any person who wishes to use these data for any form of commercial purpose must first enter into a commercial licensing and benefit sharing arrangement with the Government of Malawi. For further information, contact the Access and Benefit-sharing National Focal Point (ABS NFP) for Malawi registered with CBD at https://www.cbd.int/information/nfp.shtml.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

The Malawi cichlid samples were collected ethically under prescribed permits, and the results and data are published under an access and benefit sharing agreement with the Government of Malawi. We acknowledge the contributions of the employees of Stuart M. Grant Ltd., the Malawi Department of Fisheries, and the Government of Malawi for their assistance in the collection of samples and the generation of data and results. We also thank Jonathan Price for helpful discussions and insight pertaining to the analysis performed in this paper. F.X.Q. was supported by the Wellcome Trust (108864/B/15/Z) and the Cambridge Commonwealth European and International Trust. M.V.A. is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 101027241. M.B. is funded by a Harding Distinguished Postgraduate Scholarship. C.U.Y. was funded by the Cambridge Commonwealth European and International Trust. This work was supported by the following grants to R.D. (Wellcome 207492/z/17/z) and E.M.: Wellcome Trust Senior Investigator Award (219475/Z/19/Z) and Cancer Research UK award (C13474/A27826).

Author contributions: F.X.Q. conceived the idea, wrote the code for data analysis, and prepared the manuscript. M.V.A. and M.B. contributed to the code and participated in discussions. M.V.A., C.U.Y., and K.S. carried out the PCR validation experiments, and M.B., B.F., and B.J. were responsible for sequencing and genome assembly. R.Z., B.R., G.F.T., M.E.S., and H.S. provided fish samples and essential project resources. M.H., R.D., and E.M. supervised the project and contributed to discussions. All authors critically reviewed and made contributions to the final manuscript.

References

Albertson RC, Kawasaki KC, Tetrault ER, Powder KE. 2018. Genetic analyses in Lake Malawi cichlids identify new roles for Fgf signaling in scale shape variation. *Commun Biol* **1:** 55. doi:10.1038/s42003-018-0060-4

Almeida MV, Vernaz G, Putman ALK, Miska EA. 2022. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. *Trends Genet* **38:** 529–553. doi:10.1016/j.tig.2022.02.009

Bao W, Kojima KK, Kohany O. 2015. RepBase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6:** 11. doi:10.1186/s13100-015-0041-9

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, Simakov O, Ng AY, Lim ZW, Bezault E, et al. 2014. The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513:** 375–381. doi:10.1038/nature13726

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10:** 421. doi:10.1186/1471-2105-10-421

Carleton KL, Conte MA, Malinsky M, Nandamuri SP, Sandkam BA, Meier JI, Mwaiko S, Seehausen O, Kocher TD. 2020. Movement of transposable

elements contributes to cichlid diversity. *Mol Ecol* **29:** 4956–4969. doi:10.1111/mec.15685
Cayuela H, Dorant Y, Mérot C, Laporte M, Normandeau E, Gagnon-Harvey S, Clément M, Sirois P, Bernatchez L. 2021. Thermal adaptation rather than demographic history drives genetic structure inferred by copy number variants in a marine fish. *Mol Ecol* **30:** 1624–1641. doi:10.1111/mec.15835
The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87. doi:10.1038/nature04072
Conte MA, Joshi R, Moore EC, Nandamuri SP, Gammerdinger WJ, Roberts RB, Carleton KL, Lien S, Kocher TD. 2019. Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *GigaScience* **8:** giz030. doi:10.1093/gigascience/giz030
Crysnanto D, Leonard AS, Fang Z-H, Pausch H. 2021. Novel functional sequences uncovered through a bovine multiassembly graph. *Proc Natl Acad Sci* **118:** e2101056118. doi:10.1073/pnas.2101056118
Ebler J, Ebert P, Clarke WE, Rausch T, Audano PA, Houwaart T, Mao Y, Korbel JO, Eichler EE, Zody MC, et al. 2022. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet* **54:** 518–525. doi:10.1038/s41588-022-01043-w
Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133–138. doi:10.1126/science.1162986
Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, Chang X, Seaman JD, Rounthwaite R, Ebler J, et al. 2020. Pangenome graphs. *Annu Rev Genomics Hum Genet* **21:** 139–162. doi:10.1146/annurev-genom-120219-080406
Fan S, Meyer A. 2014. Evolution of genomic structural variation and genomic architecture in the adaptive radiations of African cichlid fishes. *Front Genet* **5:** 163. doi:10.3389/fgene.2014.00163
Fang B, Edwards SV. 2024. Fitness consequences of structural variation inferred from a House Finch pangenome. *Proc Natl Acad Sci* **121:** e2409943121. doi:10.1073/pnas.2409943121
Fryer G, Iles TD. 1972. *The cichlid fishes of the great lakes of Africa: their biology and evolution*. Oliver and Boyd, Edinburgh.
Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, et al. 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet* **51:** 1044–1051. doi:10.1038/s41588-019-0410-2
Gao Y, Yang X, Chen H, Tan X, Yang Z, Deng L, Wang B, Kong S, Li S, Cui Y, et al. 2023. A pangenome reference of 36 Chinese populations. *Nature* **619:** 112–121. doi:10.1038/s41586-023-06173-7
Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, et al. 2018. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* **36:** 875–879. doi:10.1038/nbt.4227
Gong Y, Li Y, Liu X, Ma Y, Jiang L. 2023. A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? *J Anim Sci Biotechnol* **14:** 73. doi:10.1186/s40104-023-00860-1
Groza C, Chen X, Wheeler TJ, Bourque G, Goubert C. 2024. A unified framework to analyze transposable element insertion polymorphisms using graph genomes. *Nat Commun* **15:** 8915. doi:10.1038/s41467-024-53294-2
Gu Z. 2022. Complex heatmap visualization. *iMeta* **1:** e43. doi:10.1002/imt2.43
Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. 2022. ODGI: understanding pangenome graphs. *Bioinformatics* **38:** 3319–3326. doi:10.1093/bioinformatics/btac308
Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29:** 1072–1075. doi:10.1093/bioinformatics/btt086
Hulsey CD, Meyer A, Streelman JT. 2020. Convergent evolution of cichlid fish pharyngeal jaw dentitions in mollusk-crushing predators: comparative X-ray computed tomography of tooth sizes, numbers, and replacement. *Integr Comp Biol* **60:** 656–664. doi:10.1093/icb/icaa089
Igolkina AA, Vorbrugg S, Rabanal FA, Liu H-J, Ashkenazy H, Kornienko AE, Fitz J, Collenberg M, Kubica C, Morales AM, et al. 2024. Towards an unbiased characterization of genetic polymorphism. bioRxiv doi:10.1101/2024.05.30.596703
Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. 2020. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res* **9:** ELIXIR-709. doi:10.12688/f1000research.24956.2
Konings A. 1989. *Malawi cichlids in their natural habitat*. Verduijn Cichlids and Lake Fish Movies, Herten, Germany.
Kratochwil CF, Liang Y, Urban S, Torres-Dowdall J, Meyer A. 2019. Evolutionary dynamics of structural variation at a key locus for color pattern diversification in cichlid fishes. *Genome Biol Evol* **11:** 3452–3465. doi:10.1093/gbe/evz261
Kratochwil CF, Kautt AF, Nater A, Härer A, Liang Y, Henning F, Meyer A. 2022. An intronic transposon insertion associates with a *trans*-species color polymorphism in Midas cichlid fishes. *Nat Commun* **13:** 296. doi:10.1038/s41467-021-27685-8
Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118
Lecomte L, Árnyasi M, Ferchaud A-L, Kent M, Lien S, Stenløkk K, Sylvestre F, Bernatchez L, Mérot C. 2024. Investigating structural variant, indel and single nucleotide polymorphism differentiation between locally adapted Atlantic Salmon populations. *Evol Appl* **17:** e13653. doi:10.1111/eva.13653
Li H, Feng X, Chu C. 2020. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* **21:** 265. doi:10.1186/s13059-020-02168-z
Li H, Wang S, Chai S, Yang Z, Zhang Q, Xin H, Xu Y, Lin S, Chen X, Yao Z, et al. 2022. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat Commun* **13:** 682. doi:10.1038/s41467-022-28362-0
Li R, Gong M, Zhang X, Wang F, Liu Z, Zhang L, Yang Q, Xu Y, Xu M, Zhang H, et al. 2023. A sheep pangenome reveals the spectrum of structural variations and their effects on tail phenotypes. *Genome Res* **33:** 463–477. doi:10.1101/gr.277372.122
Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617:** 312–324. doi:10.1038/s41586-023-05896-x
Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, Miska EA, Durbin R, Genner MJ, Turner GF. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350:** 1493–1498. doi:10.1126/science.aac9927
Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, Durbin R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat Ecol Evol* **2:** 1940–1955. doi:10.1038/s41559-018-0717-x
Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* **38:** 4647–4654. doi:10.1093/molbev/msab199
Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol Evol* **35:** 561–572. doi:10.1016/j.tree.2020.03.002
Mérot C, Stenløkk KSR, Venney C, Laporte M, Moser M, Normandeau E, Árnyasi M, Kent M, Rougeux C, Flynn JM, et al. 2023. Genome assembly, structural variants, and genetic differentiation between lake whitefish young species pairs (*Coregonus* sp.) with long and short reads. *Mol Ecol* **32:** 1458–1477. doi:10.1111/mec.16468
Mikheyev AS, Tin MMY. 2014. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* **14:** 1097–1102. doi:10.1111/1755-0998.12324
Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res* **49:** D412–D419. doi:10.1093/nar/gkaa913
Munby H, Linderoth T, Fischer B, Du M, Vernaz G, Tyers AM, Ngatunga BP, Shechonge A, Denise H, McCarthy SA, et al. 2021. Differential use of multiple genetic sex determination systems in divergent ecomorphs of an African crater lake cichlid. bioRxiv doi:10.1101/2021.08.05.455235
Nandamuri SP, Schulte JE, Yourick MR, Sandkam BA, Behrens KA, Schreiner MM, Dayanim M, Sweatt G, Conte MA, Juntti SA, et al. 2023. A second locus contributing to the differential expression of the blue sensitive opsin SWS2A in Lake Malawi cichlids. *Hydrobiologia* **850:** 2331–2353. doi:10.1007/s10750-022-05027-z
Ngoepe N, Muschick M, Kishe MA, Mwaiko S, Temoltzin-Loranca Y, King L, Courtney Mustaphi C, Heiri O, Wienhues G, Vogel H, et al. 2023. A continuous fish fossil record reveals key insights into adaptive radiation. *Nature* **622:** 315–320. doi:10.1038/s41586-023-06603-6
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20:** 289–290. doi:10.1093/bioinformatics/btg412
Penso-Dolfin L, Man A, Mehta T, Haerty W, Di Palma F. 2020. Analysis of structural variants in four African cichlids highlights an association with developmental and immune related genes. *BMC Evol Biol* **20:** 69. doi:10.1186/s12862-020-01629-0
Plessy C, Mansfield MJ, Bliznina A, Masunaga A, West C, Tan Y, Liu AW, Grašič J, Del Río Pisula MS, Sánchez-Serna G, et al. 2024. Extreme

genome scrambling in marine planktonic *Oikopleura dioica* cryptic species. *Genome Res* **34:** 426–440. doi:10.1101/gr.278295.123
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033
Rasko DA, Rosovitz MJ, Myers GSA, Mongodin EF, Fricke WF, Gajer P, Crabtree J, Sebaihia M, Thomson NR, Chaudhuri R, et al. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* **190:** 6881–6893. doi:10.1128/JB.00619-08
R Core Team. 2021. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592:** 737–746. doi:10.1038/s41586-021-03451-0
Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26. doi:10.1038/nbt.1754
Salzburger W. 2018. Understanding explosive diversification through cichlid fish genomics. *Nat Rev Genet* **19:** 705–717. doi:10.1038/s41576-018-0043-9
Santos ME, Braasch I, Boileau N, Meyer BS, Sauteur L, Böhne A, Belting H-G, Affolter M, Salzburger W. 2014. The evolution of cichlid fish egg-spots is linked with a *cis*-regulatory change. *Nat Commun* **5:** 5149. doi:10.1038/ncomms6149
Santos ME, Lopes JF, Kratochwil CF. 2023. East African cichlid fishes. *EvoDevo* **14:** 1. doi:10.1186/s13227-022-00205-5
Schulte JE, O'Brien CS, Conte MA, O'Quin KE, Carleton KL. 2014. Interspecific variation in *Rx1* expression controls opsin expression and causes visual system diversity in African cichlid fishes. *Mol Biol Evol* **31:** 2297–2308. doi:10.1093/molbev/msu172
Secomandi S, Gallo GR, Sozzoni M, Iannucci A, Galati E, Abueg L, Balacco J, Caprioli M, Chow W, Ciofi C, et al. 2023. A chromosome-level reference genome and pangenome for barn swallow population genomics. *Cell Rep* **42:** 111992. doi:10.1016/j.celrep.2023.111992
Sherman RM, Salzberg SL. 2020. Pan-genomics in the human genome era. *Nat Rev Genet* **21:** 243–254. doi:10.1038/s41576-020-0210-7
Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **51:** 30–35. doi:10.1038/s41588-018-0273-y
Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12:** 2. doi:10.1186/s13100-020-00230-y
Stuart KC, Edwards RJ, Sherwin WB, Rollins LA. 2023. Contrasting patterns of single nucleotide polymorphisms and structural variation across multiple invasions. *Mol Biol Evol* **40:** msad046. doi:10.1093/molbev/msad046
Svardal H, Salzburger W, Malinsky M. 2020. Genetic variation and hybridization in evolutionary radiations of cichlid fishes. *Annu Rev Anim Biosci* **9:** 55–79. doi:10.1146/annurev-animal-061220-023129
Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome." *Proc Natl Acad Sci* **102:** 13950–13955. doi:10.1073/pnas.0506758102
Vernikos G, Medini D, Riley DR, Tettelin H. 2015. Ten years of pan-genome analyses. *Curr Opin Microbiol* **23:** 148–154. doi:10.1016/j.mib.2014.11.016
Wang K, Hua G, Li J, Yang Y, Zhang C, Yang L, Hu X, Scheben A, Wu Y, Gong P, et al. 2024. Duck pan-genome reveals two transposon insertions caused bodyweight enlarging and white plumage phenotype formation during evolution. *iMeta* **3:** e154. doi:10.1002/imt2.154
Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31:** 3350–3352. doi:10.1093/bioinformatics/btv383
Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. 2019. Welcome to the tidyverse. *J Open Source Softw* **4:** 1686. doi:10.21105/joss.01686
York RA, Patil C, Abdilleh K, Johnson ZV, Conte MA, Genner MJ, McGrath PT, Fraser HB, Fernald RD, Streelman JT. 2018. Behavior-dependent *cis* regulation reveals genes and pathways associated with bower building in cichlid fishes. *Proc Natl Acad Sci* **115:** E11081–E11090. doi:10.1073/pnas.1810140115

Received June 16, 2024; accepted in revised form February 6, 2025.