Cambridge Lake Malawi cichlids VCF release March 2017

Contacts:
Milan Malinsky: mm21@sanger.ac.uk; Hannes Svardal: hs10@sanger.ac.uk
Richard Durbin: rd@sanger.ac.uk

Sample selection:
Our sample selection provides broad coverage of all the major lineages of the rapidly radiating cichlid tribe Haplochromini in Lake Malawi, with specimens from:

1. Eight species from the 'mbuna' group of mainly shallow-water rock-dwelling cichlids. Our specimens represent much of the diversity of this group, covering 6 out of 10 genera defined by Ribbink *et al.* in their detailed classification of 196 Malawian mbuna species[1,2].
2. Nine species of 'deep water' benthic haplochromines: many of these species are found at depths >50m, a 'twilight' zone with very little visible light; a few inhabit shallower water but are often crepuscular feeders, residing among rocks by day. These include members of the genera *Alticorpus* and *Aulonocara* (characterized by greatly enlarged sensory openings of their heads and lateral lines) and several of the species currently assigned to the genus *Lethrinops*.
3. Forty-one species of cichlids found predominantly in shallow waters close to the shore (like 'mbuna'), but on sandy or muddy lake floor and the transition zones between sandy and rocky habitats. This is a very diverse group of cichlids with hundreds of described species[3,4], including for example large (over 35cm) predators such as *Buccochromis nototaenia*, the invertebrate picker *Placidochromis johstoni*, molluscivores such as *Chilotilapia rhoadesii* and *Mylochromis anaphyrmus*, and cichlids that filter the sandy sediment like *Ctenopharynx nitidus*. We refer to this group collectively as 'shallow benthics'.
4. Four species of zooplankton-feeding, shoaling cichlids which are commonly referred to as 'utaka'. These species feed in the water column, but their distribution is limited to locations close to the shore[1,4]. All four sequenced 'utaka' species belong to the genus *Copadichromis.*
5. Three out of eight described species from the genus *Rhampochromis*[1,5] - large pelagic (open-water) piscivores, feeding mainly on lake sardines[4]. Cichlids are primarily bottom-dwellers, but members of this group [and *Diplotaxodon* (see below)] have undergone extensive changes in morphology and behavior to invade the large pelagic habitat in Lake Malawi.
6. Three out of seven scientifically described species of *Diplotaxodon* (ref. [1,4,6], p. 198), and three undescribed species: *D. macrops* 'black dorsal'[3,4], *D.* 'ngolube' (ref. [1,7] p. 239), and *D.* 'white back similis' (new undescribed species). Like *Rhampochromis*, *Diplotaxodon* are pelagic cichlids, but are typically found at greater depths (>50m) and their diet consists mainly of zooplankton, rather than fish. We include in this group also the species *Pallidochromis tokolosh,* which is morphologically intermediate between the two genera and is a slightly more benthic form (ref. [1,8], pp. 198-199), but its genetic affinities are more to the genus *Diplotaxodon*, as we show in this manuscript.
7. *Astatotilapia calliptera* - one of only two Lake Malawi haplochromine cichlids able to cross the lake-river barrier. *A. calliptera* is a versatile, relatively small cichlid (~10-15cm), common in the rivers throughout the Lake Malawi catchment. In Lake Malawi, *A. calliptera* frequents shallow sheltered bays with muddy sediment and aquatic plants, often feeding on snails (ref. [4,9], p. 281). It has been suggested that it may be related to the ancestral lake-river generalist species that seeded most or perhaps all of the Lake Malawi haplochromine radiation[4,6]. Therefore, we sampled *A. calliptera* genetic variation extensively, including 16 specimens (four from Lake Malawi itself and twelve more from the broader Lake Malawi catchment).

   The other species able to cross the lake-river barrier is *Serranochromis robustus*: a large predator often seen in very shallow water near river estuaries (ref. [4], p. 277) and common in rivers to the south-west of Lake Malawi, including the Zambezi system[7]. However, *S. robustus* is not a part of the Lake Malawi haplochromine radiation, but instead belongs to a distant group sometimes referred to as the 'Congo clade'[8]. For a list of all samples and details of our assignment of Lake Malawi species into these seven lineages see Supplementary Table 1.

Supplementary Table 1:

An overview of Lake Malawi and species and samples sequenced in this study. Approximate coverage and Illumina sequencing chemistry used (v3 or v4) are indicated.

Group	Species	Number of individuals: ~coverage/chemistry				Total individuals
		~15x/v3	~15x/v4	~6x/v3	~6x/v4	
mbuna	*Cynotilapia afra*		1			**1**
	Cynotilapia axelrodi		1			**1**
	Genyochromis mento		1			**1**
	Iodotropheus sprengerae		1			**1**
	Labeotropheus trewavasae				1	**1**
	Metriaclima zebra		1			**1**
	Petrotilapia genalutea		1			**1**
	Tropheops tropheops		1			**1**
deep benthic	*Alticorpus geoffreyi*	1				**1**
	Alticorpus macrocleithrum	1				**1**
	Aulonocara 'minutus'		1			**1**
	Aulonocara steveni	1				**1**
	Aulonocara stuartgranti	3				**3 (trio)**
	Aulonocara 'yellow'		1			**1**
	Lethrinops gossei	1				**1**
	Lethrinops longimanus 'redhead'	1				**1**
	Lethrinops 'oliveri'	1				**1**
shallow benthic	*Buccochromis nototaenia*		1			**1**
	Buccochromis rhoadesii	1				**1**
	Champsochromis caeruleus		2			**2**
	Chilotilapia rhoadesii		1		3	**4**
	Copadichromis cf trewavasae	1				**1**
	Ctenopharynx intermedius		1		2	**3**
	Ctenopharynx nitidus		2			**2**
	Dimidiochromis compressiceps		1			**1**
	Dimidiochromis dimidiatus		1			**1**
	Dimidiochromis kiwinge		1			**1**
	Dimidiochromis strigatus		2			**2**
	Fossorochromis rostratus				2	**2**
	Hemitaeniochromis spilopterus		2			**2**
	Hemitilapia oxyrhynchus		2			**2**
	Lethrinops albus	1				**1**
	Lethrinops auritus	1				**1**
	Lethrinops lethrinus	4				**4 (trio+1)**
	Mylochromis anaphyrmus	1		4		**5**
	Mylochromis ericotaenia		1			**1**
	Mylochromis melanotaenia		1			**1**
	Nimbochromis linni	1				**1**
	Nimbochromis livingstoni	1				**1**
	Nimbochromis polystigma	1				**1**
	Otopharynx brooksi 'nkhata'	1				**1**
	Otopharynx lithobates		1			**1**
	Otopharynx speciosus		2			**2**
	Otopharynx tetrastigma (Lake Ilamba)	1				**1**
	Placidochromis electra	1				**1**
	Placidochromis johnstoni	1				**1**
	Placidochromis cf. longimanus		1		4	**5**
	Placidochromis milomo	1				**1**
	Placidochromis subocularis				8	**8**
	Protomelas ornatus		2			**2**
	Stigmatochromis guttatus		1			**1**
	Stigmatochromis modestus		1			**1**
	Taeniochromis holotaenia		1			**1**
	Taeniolethrinops furcicauda		1			**1**
	Taeniolethrinops macrorhynchus		1			**1**
	Taeniolethrinops praeorbitalis		1			**1**
	Tremitochranus placodon	1		4		**5**
	Tyrannochromis nigriventer		1			**1**
utaka	*Copadichromis likomae*		1			**1**
	Copadichromis quadrimaculatus	1				**1**
	Copadichromis trimaculatus		1			**1**
	Copadichromis virginalis	1			4	**5**
Rhampochromis	*Rhamphochromis esox*		1			**1**
	Rhamphochromis longiceps		1			**1**
	Rhamphochromis woodi		1			**1**
Diplotaxodon	*Diplotaxodon greenwoodi*		1			**1**
	Diplotaxodon limnothrissa		1			**1**
	Diplotaxodon macrops		1			**1**
	Diplotaxodon 'macrops black dorsal'		1			**1**
	Diplotaxodon 'macrops ngulube'		1			**1**
	Diplotaxodon 'similis white back'		1			**1**
	Pallidochromis tokolosh		1			**1**
A. calliptera	*Astatotilapia calliptera*	8	13			**21 (trio+18)**
	Number of species: 73	**36**	**67**	**8**	**23**	**134**

Supplementary Table 2:

Individual BioSample accessions for whole genome sequencing data used in this study.

Samples	**BioSample Accessions**
Lake Malawi populations	SAMEA1877409, SAMEA1877411, SAMEA1877414, SAMEA1877417, SAMEA1877421, SAMEA1877429, SAMEA1877440, SAMEA1877451, SAMEA1877455, SAMEA1877459, SAMEA1877464, SAMEA1877472, SAMEA1877476, SAMEA1877480, SAMEA1877484, SAMEA1877499, SAMEA1877503, SAMEA1904322, SAMEA1904324, SAMEA1904329-SAMEA1904331, SAMEA2661216-SAMEA2661246, SAMEA2661250-SAMEA2661253, SAMEA2661255-SAMEA2661258, SAMEA2661260, SAMEA2661262, SAMEA2661264-SAMEA2661270, SAMEA2661272, SAMEA2661275-SAMEA2661282, SAMEA2661287-SAMEA2661290, SAMEA3388853-SAMEA3388860, SAMEA3388862-SAMEA3388864, SAMEA3388868, SAMEA3388870-SAMEA3388874
Lake Malawi trios	SAMEA1920096-SAMEA1920098 *A. stuartgranti* SAMEA1920093-SAMEA1920095 *L. lethrinus* SAMEA1920090-SAMEA1920092 *A. calliptera* Salima, Lake Malawi
***A. calliptera* Malawi catchment**	SAMEA1877400, SAMEA1904326, SAMEA1904327, SAMEA2661273(currently misannotated on NCBI, this sample is *A. calliptera* Kingiri), SAMEA2661381-SAMEA2661385, SAMEA2661389-SAMEA2661391
***A. calliptera* Indian Ocean catchments**	SAMEA1904323, SAMEA1904328, SAMEA2661386-SAMEA2661388

Methods:

DNA extraction and sequencing:

DNA was extracted from fin clips using the PureLink® Genomic DNA extraction kit (Life Technologies). Genomic libraries for paired-end sequencing on the Illumina HiSeq 2000 machine were prepared according to the Illumina TruSeq HT protocol to obtain paired-end reads with mean insert size of 300-500bp. As detailed in Supplementary Table 1, we used either Illumina HiSeq v3 chemistry (generating 100bp paired-end reads) or Illumina HiSeq v4 reagents (125bp paired-end reads). Low coverage (~6x) samples with v4 reagents were multiplexed 12 per lane. High coverage (~15x) v4 samples were multiplexed four per lane. For high coverage (~15x) v3 samples, a multiplexed library with 8 samples was sequenced over three lanes. The nine trio samples were multiplexed across eight lanes using the v3 chemistry, delivering approximately 40x coverage per individual. Raw data have been deposited at the NCBI Sequence Read Archive under BioProject PRJEB1254; sample accessions are listed in Supplementary Table 2.

Alignment:

Reads were aligned to the *Metriaclima zebra* reference assembly version 1.1-prescreen (http://www.broadinstitute.org/ftp/pub/assemblies/fish/M_zebra/MetZeb1.1_prescreen/M_zebra_v0.assembly.fasta) using the `bwa-mem v.0.7.10` algorithm with default options. For the trio samples, we aligned data from only three of the eight lanes, aiming to have the same genome coverage for the trios as for the remaining (15x) samples. For each sample, 96-98% of reads could be aligned to the reference. Duplicate reads were marked on both per-lane and per sample basis using the `MarkDuplicates` tool from the `Picard` software package with default options (http://broadinstitute.github.io/picard). Local realignment around indels was performed on both per lane and per sample basis using the `IndelRealigner` tool from the GATK v.3.3.0 software package.

Variant calling, filtering, and genotype refinement:

Briefly, SNP and short indel variants against the *M. zebra* reference were called independently using GATK v3.3.0 haplotype caller and samtools/bcftools v.1.1. Variant filtering was then performed on the GATK variant calls using hard filters based on overall depth, quality by depth, excess missingness, excess of reads with zero mapping quality, strand/mapping bias, and inbreeding coefficient (see below). After filtering the GATK dataset, we performed an intersection of GATK and samtools sites and kept only variant sites present in both datasets. If the GATK and samtools alleles differed at a particular locus, we kept the GATK allele. At this point, multiallelic sites were excluded and we used genotype likelihoods output by GATK sites to perform genotype refinement, imputation, and phasing in BEAGLE v.4.0. Indels were retained for the genotype refinement step, but later generally excluded from analyses using `vcftools v0.1.12b` option `--remove-indels`, except where specifically indicated.

The particular commands/parameters used were:
GATK haplotype caller (per sample):

```
java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R REFERENCE.fa --
emitRefConfidence GVCF --variant_index_type LINEAR --variant_index_parameter
128000 -I SAMPLEn.bam -o GATK_SAMPLEn.g.vcf
```

Haplotype caller per-sample files were combined using the GATK `GenotypeGVCFs` tool with the `--includeNonVariantSites` option so that every basepair of the assembly was represented in the multisample VCF file
Hard filters were applied to the following GATK annotations:

```
Minimal inbreeding coefficient: 'InbreedingCoeff < -0.6'
Minimum overall read depth: 'DP < 1000'
Maximum overall read depth: 'DP > 3000' (except for mtDNA: scaffolds 747,2036)
```

```
Max phred-scaled p-value from Fisher's exact test to detect strand bias: 'FS >
40.0' (except for mtDNA: scaffolds 747,2036)
QualityByDepth: 'QD < 2.0’
Excess Missingness: 'NCC > 32' (>16 individuals with missing data)
More than 10% of reads have mapping quality zero: '(MQ0/(1.0*DP)) > 0.10'
Low mapping quality: 'MQ < 40.0'
```

samtools calling (multisample):

```
samtools mpileup -t DP,DPR,INFO/DPR -C50 -pm2 -F0.2 –ugf REFERENCE.fa SAMPLE1.bam
SAMPLE2.bam ... | bcftools call -vmO z -f GQ -o samtools_VARIANTS.vcf.gz
```

The consensus GATK and samtools call set was obtained using the `GATK`:

```
java -Xmx10000m -jar GenomeAnalysisTK.jar -T SelectVariants -R REFERENCE.fa --
variant onlyVariants_filtered.vcf.gz -o onlyVariants_filtered_concord.vcf.gz --
concordance samtools_unfiltered.vcf.gz
```

BEAGLE genotype refinement (per scaffold) - specifying the trio relationships:

```
java   -jar   beagle.r1398.jar   gl=onlyVariants_filtered_concord_sc${sc}.vcf.gz
ped=../Malawi_trios.pedind nthreads=8 ibd=true ibdtrim=200 phase-its=8 impute-
its=8 out=beagle_onlyVariants_filtered_concord_sc_${sc}
```

Terms:

This is a public release in advance of publication of our genome-wide analysis of this dataset. It can be used without any restrictions for analyses concerning specific genomic region(s). If a genome-wide analysis forms a major part of a manuscript to be submitted before we publish on this dataset, we require the authors to contact us to discuss how to recognize our contribution in generating, preserving and sharing this resource.

References:

1. *The cichlid diversity of Lake Malawi/Nyasa/Niassa*. (Cichlid Press, 2004).
2. Ribbink, A. J., Marsh, B. A., Marsh, A. C., Ribbink, A. C. & Sharp, B. J. A preliminary survey of the cichlid fishes of rocky habitats in Lake Malawi. *S. Afr. J. Zool.* **18,** 149–310 (1983).
3. Genner, M. J. *et al.* Reproductive isolation among deep-water cichlid fishes of Lake Malawi differing in monochromatic male breeding dress. *Mol Ecol* **16,** 651–662 (2007).
4. Konings, A. *Malaŵi Cichlids in Their Natural Habitat*. (Cichlid Press, 2007).
5. Eccles, D. H. & Trewavas, E. *Malawian Cichlid Fishes*. (Lake Fish Movies, 1989).
6. Joyce, D. A. *et al.* Repeated colonization and hybridization in Lake Malawi cichlids. *Curr. Biol.* **21,** R108–9 (2011).
7. Joyce, D. A. *et al.* An extant cichlid fish radiation emerged in an extinct Pleistocene lake. *Nature* **435,** 90–95 (2005).
8. Schwarzer, J. *et al.* Repeated trans-watershed hybridization among haplochromine cichlids (Cichlidae) was triggered by Neogene landscape evolution. *Proceedings of the Royal Society B: Biological Sciences* **279,** 4389–4398 (2012).
9. Brawand, D. *et al.* The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513,** 375–381 (2014).