Chapter 3

New genome sequence datasets

3.1 Introduction

3.1.1 Overview of whole-genome data

The genome assemblies and annotations from the Cichlid Genome Consortium provide a key resource facilitating in-depth studies of the individual radiations in the Great Lakes of East Africa. Taking advantage of this opportunity, we obtained almost 3,000 Gbp of raw whole genome sequence data from 281 individuals covering 88 species from in and around Lake Malawi (Table 3.1). This rich dataset has facilitated all the cichlid work described in the rest of this thesis.

Table 3.1 **Whole genome sequencing at Sanger Institute - an overview.**

Sample type	Number of individuals	Primary goals	Coverage per individual	Overall amount of sequence
Mother - father - offspring trio	9 (three trios)	De-novo assembly Haplotype phasing	~40X	~360Gbp
Lake Malawi HC	83	Assessing between species diversity	~15X	~1,245Gbp
Lake Malawi LC	34	Assessing within species diversity	~5X	~170Gbp
Astatotilapia calliptera populations	19	Diversity in rivers and lakes around Lake Malawi	~15X	~285Gbp
Crater lake *Astatotilapia*	130	Detailed view of divergence between incipient species	~15X - n:12 ~5X - n:118	~180Gbp ~590Gbp
Other	6	Various; see text	~15X - n:5 ~60X - n:1	~75Gbp ~60Gbp
Total:	281			~2,965Gbp

The 281 cichlid samples were obtained from our collaborators on this project George Turner (University of Bangor) and Martin Genner (University of Bristol), who have also assigned individual specimens to species. Sequencing progressed in three phases:

a first batch of samples was sequenced in Spring 2013, second batch in Autumn 2014, and the third in Summer 2015. Tables 3.2 and 3.3 describe all the individual specimens sequenced and their assignment to sequencing batches. Samples from the Summer 2015 batch are not included in any analysis because they were sequenced only a few weeks before the completion of this thesis.

Lake Malawi harbours five species of tilapiine cichlids (tribe Tilapiini): four *Oreochromis* species and *Tilapia rendalli.* These fish, although present in Lake Malawi, do not form a part of the main Lake Malawi adaptive radiation and represent a separate evolutionary lineage. Therefore, they were not included in our sampling.

We have focussed on the rapidly radiating tribe Haplochromini. Our Lake Malawi samples (Table 3.2) cover all its major haplochromine lineages, including specimens from the following groups:

1. Nine species from the ‘mbuna’ group of mainly shallow-water rock-dwelling cichlids. Our specimens represent much of the diversity of this group, covering 7 out of 10 genera defined by Ribbink *et al.* [124]. Mbuna is a common name given to these fish by the Tonga people of Malawi [124]. Since the Ribbink *et al.* detailed classification of 196 mbuna species [124] found in the Malawian waters, dozens new species have been described [125], members of the *Pseudotropheus tropheops* species-complex have been assigned to a new (sub)genus *Tropheops* [126], and members of the *Pseudotropheus zebra* species-complex have been assigned to a new genus *Metriaclima* [127].
2. Ten species of pelagic (open-water) cichlids. Cichlids are primarily bottom-dwellers, but members of three genera (*Diplotaxodon*, *Rhampochromis*, and *Pallidochromis*) have undergone extensive changes in morphology and behaviour to become pelagic piscivores (ecotype h in Figure 2.1), feeding mainly on crustaceans and lake sardines [125]. We have sequenced three out of seven scientifically described species of *Diplotaxodon* [87, p. 198], and three undescribed species: *D.* ‘macrops black dorsal’ [128], *D.* ‘ngolube’ [87, p. 239], and *D.* ‘white back similis’ (M. Genner, pers. comm.). There are eight described species of *Rhampochromis* [129], or which we have sampled three. Finally, *Pallidochromis* is a single species genus - *P. tokolosh* is morphologically intermediate between the other two genera and is a slightly more benthic form [87, pp. 198-199].
3. Twelve species of deep water benthic haplochromines, generally caught at depths of 50m, a ‘twilight’ zone with very little visible light. These include members of the genera *Alticorpus* and *Aulonocara*, characterised by greatly enlarged sensory openings of their heads and lateral lines, several of the species currently assigned

to the genus *Lethrinops*, and *Otopharynx speciosus.* There are also 47 (described and undescribed) deep water species of *Placidochromis* [87, pp. 104-197]. and three or four deep water species of *Stigmatochromis* [125, pp. 405-408]. However, we have not obtained any deep-water specimens from these two genera.

4. Forty-five 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 [125], including for example large (over 35cm) predators such as *Buccochromis nototaenia*, the small plankton-feeding shoaling *Copadichromis*, and mollucivores such as *Chilotilapia rhoadesii* and *Mylochromis anaphyrmus.* We refer to this group as 'sand dwellers'.
5. Two haplochromine cichlids found in Lake Malawi are able to cross the lake-river barrier: *Astatotilapia calliptera* and *Serranochromis robustus.* A versatile, relatively small cichlid (~10-15cm) common in the rivers throughout Lake Malawi catchment, *A. calliptera* in Lake Malawi frequents shallow sheltered bays with muddy sediment and aquatic plants, often feeding on snails [125, 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 [86, 125]. For this and other reasons to be discussed later, we sampled *A. calliptera* genetic variation extensively. *S. robustus* is a large predator often seen in very shallow water near river estuaries [125, p. 277] and is a common species in rivers to the south-west of Lake Malawi, including the Zambezi system [130]. Eccles and Trewavas [129, pp. 24-26] suggest that some Lake Malawi genera, especially among the larger sand-dwellers, may have ancestors allied to *S. robustus* or other members of an ancestral group of riverine species of the Zambezi system.

Table 3.2 **Lake Malawi sequencing.** Colours indicate sequencing batches: blue - Spring 2013, brown - Autumn 2014, green - Summer 2015. Symbols indicate common species groups: ∗ - 'mbuna', ● - open-water (pelagic) cichlids, ▼ - deep water sand-dwellers, ▲ - shallow water sand-dwellers, □ - lake-river generalists

Panel A: Population genomics samples

Species	Samples		Species	Samples	
	15X	5X		15X	5X
Alticorpus geoffreyi▼	1	0	*Hemitilapia oxyrhynchus*▲	1	0
Alticorpus macrocleithrum▼	1	0	*Iodotropheus sprengerae*⋆	1	0
Alticorpus peterdaviesi▼	1	0	*Labeotropheus trewavasae*⋆	1	0
Aulonocara 'blue chilumba'▼	1	0	*Lethrinops albus*▲	1	0
Aulonocara 'gold'▼	1	0	*Lethrinops auritus*▲	1	0
Aulonocara 'minutus'▼	1	0	*Lethrinops gossei*▼	1	0
Aulonocara 'yellow'▼	1	0	*Lethrinops lethrinus*▲	1	0
Aulonocara steveni▲	1	0	*Lethrinops 'longimanus redhead'*▼	1	0
Aulonocara stuartgranti 'maisoni'▲	1	0	*Lethrinops 'oliveri'*▼	1	0
Buccochromis nototaenia▲	1	0	*Metriaclima zebra*⋆	1	0
Buccochromis rhoadesii▲	1	0	*Mylochromis anaphyrmus*▲	1	4
Champsochromis caeruelus▲	2	0	*Mylochromis ericotaenia*▲	1	0
Chilotilapia rhoadesii▲	1	3	*Mylochromis melanotaenia*▲	1	0
Copadichromis borleyi▲	1	0	*Nimbochromis linni*▲	1	0
Copadichromis likomae▲	1	0	*Nimbochromis livingstoni*▲	1	0
Copadichromis mloto▲	1	0	*Nimbochromis polystigma*▲	1	0
Copadichromis quadrimaculatus▲	1	0	*Otopharynx 'brooksi nkhata'*▼	1	0
Copadichromis trimaculatus▲	1	0	*Otopharynx lithobates*▲	1	0
Copadichromis virginalis▲	1	4+2[1]	*Otopharynx speciosus*▼	2	0
Ctenopharynx intermedius▲	1	2	*Pallidochromis tokolosh*●	1	0
Ctenopharynx nitidus▲	1	0	*Petrotilapia genalutea*⋆	1	0
Ctenopharynx nitidus▲	1	0	*Placidochromis electra*▲	1	0
Cyathochromis obliquidens⋆	1	0	*Placidochromis johnstoni*▲	1	0
Cynotilapia afra⋆	1	0	*Placidochromis longimanus*▲	1	4
Cynotilapia axelrodi⋆	1	0	*Placidochromis milomo*▲	1	0
Dimidiochromis compressiceps▲	1	0	*Placidochromis subocularis*▲	0	8
Dimidiochromis dimidiatus▲	1	0	*Protomelas ornatus*▲	2	0
Dimidiochromis kiwinge▲	1	0	*Rhamphochromis esox*●	1	0
Dimidiochromis strigatus▲	1	0	*Rhamphochromis longiceps*●	1	0
Dimidiochromis strigatus▲	1	0	*Rhamphochromis woodi*●	1	0
Diplotaxodon 'ngulube'●	1	0	*Serranochromis robustus*□	1	0
Diplotaxodon 'white back similis'●	1	0	*Stigmatochromis 'guttatus'*▲	1	0
Diplotaxodon greenwoodi●	1	0	*Stigmatochromis modestus*▲	1	0
Diplotaxodon limnothrissa●	1	0	*Taeniochromis holotaenia*▲	1	0
Diplotaxodon macrops●	1	0	*Taeniolethrinops furcicauda*▲	1	0
Diplotaxodon 'macrops black dorsal'●	1	0	*Taeniolethrinops macrorhynchus*▲	1	0
Fossorochromis rostratus▲	1	3	*Taeniolethrinops praeorbitalis*▲	1	0
Genyochromis mento⋆	1	0	*Tremitochranus placodon*▲	1	4
Hemitaeniochromis spilopterus▲	1	0	*Tropheops tropheops*⋆	1	0
Hemitaeniochromis spilopterus▲	1	0	*Tyrannochromis nigriventer*▲	1	0
Hemitilapia oxyrhynchus▲	1	0			

Panel B: Deep coverage samples for genome assembly

Species	Sampling location	Relationship	Coverage	
			paired-end	mate-pair
Astatotilapia calliptera□	Salima region	father	~40X	0
Astatotilapia calliptera□	Salima region	mother	~40X	0
Astatotilapia calliptera□	Salima region	offspring	~40X	~5X
Aulonocara stuartgranti▲	Usisya region	father	~40X	0
Aulonocara stuartgranti▲	Usisya region	mother	~40X	0
Aulonocara stuartgranti▲	Usisya region	offspring	~40X	~5X
Lethrinops lethrinus▲	Mazinzi reef	father	~40X	0
Lethrinops lethrinus▲	Mazinzi reef	mother	~40X	0
Lethrinops lethrinus▲	Mazinzi reef	offspring	~40X	~5X

[1] from Lake Malombe

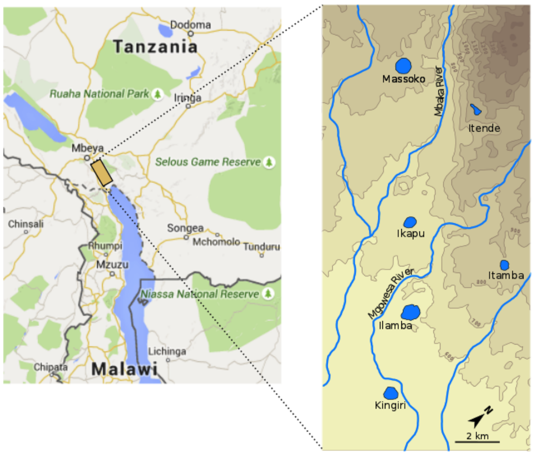


Fig. 3.1 **Map of the crater lake region in Southern Tanzania**

In the summer of 2011, Martin Genner and George Turner explored the cichlid fish fauna of crater lakes in the Rungwe District of Tanzania, approximately 40km north of Lake Malawi (Figure 3.1). They discovered haplochromine cichlids derived from *Astatotilapia calliptera* in six of the lakes, and a pair of incipient species forming within one - Lake Massoko. We have obtained whole genome sequence data from 100 *Astatotilapia* individuals from Lake Massoko, 30 individuals from Lake Itamba, one from Lake Kingiri, four from the Itupi stream - the closest water body upstream of Lake Massoko, and one from the Mbaka river - a major river downstream from Lake Massoko. Furthermore, to explore the geographical context of the crater-lake radiation and given the potential importance of *A. calliptera* ancestors in the Lake Malawi radiation, we have added 13 more *A. calliptera* from the wider Lake Malawi catchment from locations shown in Figure 3.2 (also see Table 3.3 - Panel A). The *A. calliptera* radiation in the crater lakes of Tanzanian Rungwe District is explored in detail in chapter 6.

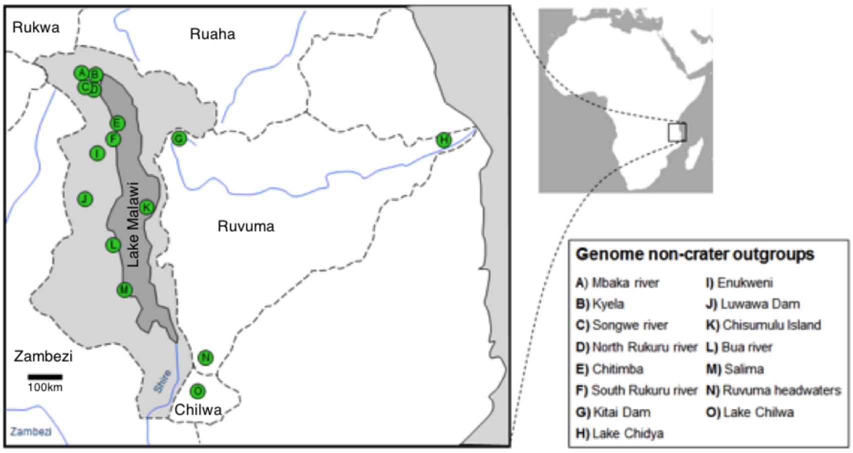


Fig. 3.2 **Collection sites of non-crater-lake *Astatotilapia calliptera* specimens.** Dotted lines represent catchment boundaries, with the Lake Malawi catchment shaded in grey. Figure from Martin Genner.

Table 3.3 **Cichlid samples from outside Lake Malawi.** Colours indicate sequencing batches: blue - Spring 2013, brown - Autumn 2014, green - Summer 2015.

Panel A: An overview of *Astatotilapia calliptera* and crater lake samples

Sampling location (ecomorph)	Samples		Latitude S	Longitude E
	15X	5X		
Lake Massoko (benthic)	6	32	9°20′00	33°45′18
Lake Massoko (littoral)	6	25	9°20′00	33°45′18
Lake Massoko (small unassigned)	0	31	9°20′00	33°45′18
Itupi stream	4	0	9°19′47	33°44′40
Lake Itamba	0	30	9°21′04	33°50′39
Lake Kingiri	1	0	9°25′00	33°51′00
Chitimba	1	0	10°34′37	34°10′14
North Rukuru	1	0	9°55′01	33°55′39
Songwe River	1	0	9°35′14	33°46′10
South Rukuru	1	0	10°45′42	34°07′33
Enukweni	1	0	11°11′14	33°52′52
Lake Chidya	1	0	10°35′49	40°09′19
Kitai Dam	1	0	10°42′22	35°11′46
Ruvuma river	1	0	14°22′22	35°32′54
Near Kyela	1	0	9°33′05	33°53′11
Luwawa Dam	1	0	12°06′57	33°43′23
Bua	1	0	13°18′30	33°32′51
Chisumulu island	1	0	~12°00′00	~34°37′00
Mbaka River	1	0	9°20′27	33°47′04
Lake Chilwa	1	0	15°22′15	35°35′30

Panel B: Other Sanger Institute sequenced samples from outside Lake Malawi

Species	Sampling location	Coverage	Notes
Astatotilapia ‘rujewa’	Ruaha river	15X	*A.* ‘rujewa’ = *A.* ‘ruaha’, a taxon discovered in 2012 in the Ruaha river in Tanzania
Astatotilapia tweddlei	Lake Chiuta	15X	*A. tweddlei* is a common species to the East of Lake Malawi
Otopharynx tetrastigma	Lake Ilamba	15X	*O. tetrastigma* is a Lake Malawi endemic, but this sample is from a similar looking species found in Lake Ilamba
Rhamphochromis ‘kingiri brevis’	Lake Kingiri	15X	Similar to *R. brevis* of Lake Malawi
Rhamphochromis ‘kingiri dwarf’	Lake Kingiri	15X	Similar to a small *Rhamphochromis* species found in Lake Chilingali
Andinoacara coeruleopunctatus	Panama	60X	A cichlid from Central America, sequenced to provide a *de novo* genome assembly as an outgroup to African cichlids

Understanding the evolutionary origins of the Lake Malawi haplochromine radiation would facilitate discussion about its early stages and tests to distinguish selection on standing variation from adaptation driven by novel genetic variants arising within the lake. A study by Joyce *et al.* (2011) is inconclusive as to whether *A. calliptera* is an outgroup or a member of the Lake Malawi flock [86]. In search for a sister species, we have sequenced two candidate species: *Astatotilapia* ‘rujewa’ and *Astatotilapia tweddlei* (Table 3.3 - Panel B). In a more recent study, Genner *et al.* claim on the basis of mitochondrial DNA sequence that *Astatotilapia* ‘rujewa’ is “immediate sister taxon” to the Lake Malawi flock [131].

In the crater lake Kingiri, Genner and Turner also discovered a pair of species of *Rhampochromis* (Table 3.3 - Panel B). The two forms have dramatically different morphology but share mtDNA haplotypes (M. Genner, pers. comm). In conjunction with the sequencing of *Rhampochromis* from Lake Malawi, the nuclear DNA of the Kingiri samples will enable us to test if the two forms have invaded Kingiri independently or if they have diverged from a common ancestor inside Lake Kingiri, representing another case of sympatric speciation in addition to Lake Massoko *Astatotilapia*.

We have also obtained data from a haplochromine species of the crater lake Ilamba whose morphology is reminiscent of *Otopharynx tetrastigma* (Table 3.3 - Panel B). The data will enable us to investigate the origin of this species.

The Cichlid Genome Project did not provide a reference genome for non-African cichlids. Furthermore, at the time of writing reference genomes are not available for any non-cichlid members of the suborder Labrodei. The closest available genome is that of medaka, sharing common ancestor with cichlids well over a hundred million years ago [93]. We have obtained high coverage data for *Andinoacara coeruleopunctatus*, a Central American cichlid (Table 3.3 - Panel B), with the aim of generating a draft reference genome, which would enable us to address questions relating to the origin of the African radiation.

3.2 Alignment, variant calling, filtering, and genotype refinement

3.2.1 DNA extraction and sequencing

I extracted DNA from fin clips using PureLink® Genomic DNA extraction kit (Life Technologies). Genomic libraries for paired-end sequencing on the Illumina HiSeq 2000 machine were prepared by the Sanger Institute sequencing core teams according to the Illumina TruSeq HT protocol to obtain 100bp (Spring 2013 batch) and 125bp (Autumn 2014 and Summer 2015 batches) *paired-end* reads. In paired-end sequencing, both ends of a DNA fragment are read - e.g. the first 100bp and the last 100bp of a 300bp fragment. The mean fragment length (also called ‘insert size’) for paired-end sequencing has been 300-500bp. Three special *mate-pair* libraries with large insert sizes (~2,000bp) were generated by the Sanger Institute’s Illumina Bespoke team to support *de novo* genome assemblies, as indicated in Table 3.2 - Panel B.

3.2.2 Alignment

Reads from samples of *A. calliptera*, *A. stuartgranti*, and *L. lethrinus* which were sequenced to ~40X coverage for genome assembly were down-sampled for studies relying on alignment to a reference genome to ~15X coverage and then processed identically to other samples.

All reads were aligned to *Metriaclima zebra* reference genome [93] using the `bwa mem v0.7.10` algorithm [132] using default options. 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`) and local realignment around indels performed on both per lane and per sample basis using the `IndelRealigner` tool from the `GATK v3.3.0` software package [133].

3.2.3 Sample call-sets

Samples were divided into two partially overlapping sets:

1. **Crater lake set:** comprising all *A. calliptera* (crater lake and all other)
2. **Lake Malawi set:** comprising all Lake Malawi samples, crater lake *Rhampochromis* and *O. tetrastigma*, and all *Astatotilapia* except from crater Lakes Massoko and Itamba

Differences against the reference genome (variants) were determined (called) independently for these two sets.

3.2.4 Variant calling, filtering, genotype refinement, and haplotype phasing

Briefly, SNP and short indel variants against the *M. zebra* reference were called independently using `GATK v3.3.0` haplotype caller [134] and `samtools/bcftools v1.1` [135]. Variant filtering was then performed on each set of variants separately using hard filters based on overall depth, overall quality score, strand/mapping bias, and inbreeding coefficient (see below). Multiallelic sites were excluded. After filtering, I selected consensus sites (i.e. performed intersection of GATK and samtools sites). At a particular locus, if the GATK and samtools alleles differed, I kept the GATK allele. Finally, I used genotype likelihoods output by GATK at consensus sites to perform genotype refinement, imputation, and phasing in `BEAGLE v.4.0` [136]. The output of this process were filtered variants and phased genotypes against the *M. zebra* reference in the `VCF` format[2].

[2] Specification is available here: `https://github.com/samtools/hts-specs`

The particular commands and parameters used were:

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
```

GATK haplotype caller (per sample), later combined using GATK's `GenotypeGVCFs` tool:

```
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
```

Hard filters applied to both datasets:

```
    Minimal inbreeding coefficient:  -0.05

    Minimum overall read depth:  600

    Maximum overall read depth:  1700 (except for mtDNA: scaffolds 747,2036)
```

Hard filters applied to the GATK dataset:

```
    Maximum phred-scaled p-value using Fisher's exact test to detect strand bias:
    20 (except for mtDNA: scaffolds 747,2036)

    Minimum accepted variant quality score:  300
```

Hard filters applied to the samtools dataset:

```
    Minimum p-value for Mann-Whitney U test of Mapping Quality vs.  Strand Bias:
    0.0001 (except for mtDNA - scaffolds 747,2036)

    Minimum accepted variant quality score:  30
```

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

```
bcftools isec -c indels -O z GATK_filtered_calls.vcf.gz samtools_filtered_calls.vcf.gz
-p GATK_samtools_intersect/
```

BEAGLE genotype refinement (per scaffold):

```
java -jar beagle.r1398.jar gl=GATK_samtools_consensus.vcf.gz phase-its=8 impute-its=8
out=beagle_GATK_sam_consensus
```

3.3 Coverage and cross-contamination estimates from data

As a part of preliminary quality control I used the `verifyBamID v1.0` [137] software to check for cross-contamination (whether the reads are contaminated as a mixture of two samples) and also to estimate genome coverage over filtered variant sites.

Cross-contamination can arise for example because multiple samples are processed at the same time and a small amount of tissue or extracted DNA is physically carried over from one sample to another, or when multiple samples are sequenced together on the same sequencing lane (this is known as 'multiplexing') and the sequencing machine is unable to decode correctly the 'tag' (short DNA sequence) that distinguishes the samples.

To check for cross-contamination, the software compares the original reads with the final variant calls and then: "Using a mathematical model that relates observed sequence reads to an hypothetical true genotype, verifyBamID tries to decide whether sequence reads match a particular individual or are more likely to be contaminated (including a small proportion of foreign DNA)".

Samples with `verifyBamID` estimated contamination >3% had considerable proportions of erroneously called variants in previous human whole-genome sequencing studies (R. Durbin, pers. comm.), so we excluded such samples from further analysis.

3.3.1 Lake Malawi call set

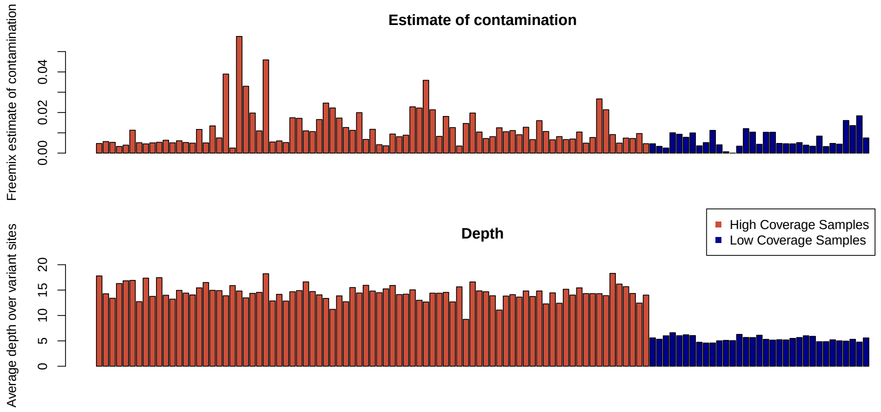


Fig. 3.3 **Cross-contamination and read-depth estimates for the Lake Malawi variant call-set.** Top: `verifyBamID` cross-contamination estimates. Bottom: read-depth over filtered variant sites.

The `verifyBamID` results for the Lake Malawi set of samples (as defined in Section 3.2.3) are shown in Figure 3.3. There are five high coverage (~15X) samples with estimated contamination scores >3%: *Aulonocara* 'blue chilumba' 5.7%, *Aulonocara stuartgranti* 'maisoni' 4.6%, *Alticorpus peterdaviesi* 3.9%, *Fossorochromis rostratus* 3.6%, *Aulonocara* 'gold' 3.2%. I eliminated these samples from all downstream analyses.

Direct estimates of read depth over filtered sites revealed that all samples were sequenced approximately to the intended coverage.

3.3.2 Crater lake call set

The `verifyBamID` results for the Crater lake set of samples (as defined in Section 3.2.3) are shown in Figure 3.4. Two samples have estimated contamination scores >3%: one from Lake Itamba 3.33% and one benthic individual from Lake Massoko 4.49%. I eliminated these two samples from all downstream analyses.

Direct estimates of read depth over filtered sites revealed that all samples were sequenced approximately to the intended coverage.

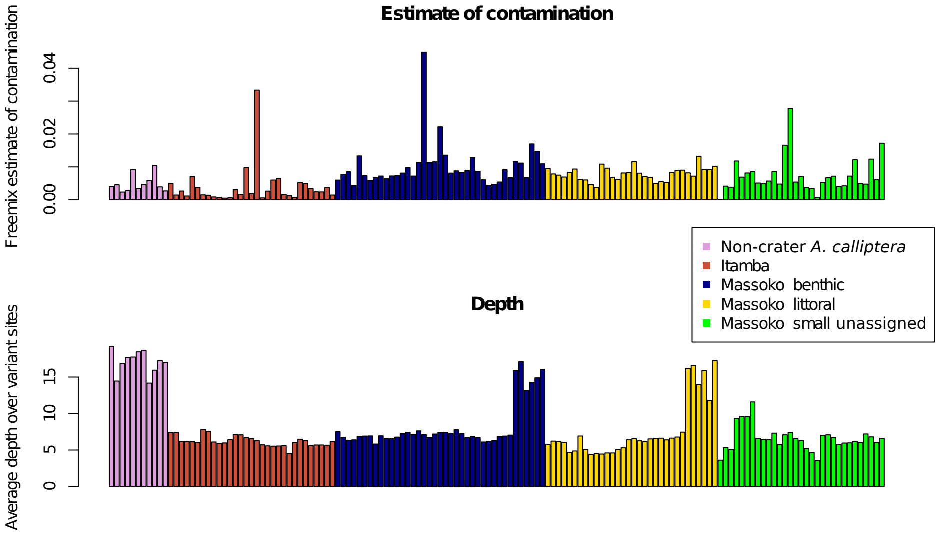


Fig. 3.4 **Cross-contamination and read-depth estimates for the Crater lake variant call-set.** Top: `verifyBamID` cross-contamination estimates. Bottom: read-depth over filtered variant sites.

3.4 Whole genome alignments

Overview

I generated a number of pairwise and multiple whole-genome alignments, following the 'UCSC paradigm' [138]. First I generated all possible pairwise alignments between the assemblies generated by the Cichlid Genome Consortium: the Lake Malawi *M. zebra* reference genome, the Lake Victoria *P. nyererei*, Lake Tanganyika *A. burtoni* and *N. brichardi*, and the Nile Tilapia *O. niloticus.* Then I added pairwise alignments between the genomes of these cichlids and the reference genomes of three other teleost species (medaka, stickleback, and zebrafish). Finally, I generated new contiguous genome-wide multiple alignments of these eight species in *M. zebra*, *P. nyererei*, *A. burtoni*, and *N. brichardi* genomic coordinates. A multiple alignment in *O. niloticus* coordinates is available from the Cichlid Genome Consortium [93].

Applications

Apart from being useful in their own right for studies of sequence evolution between the five cichlid species included, the alignments facilitate important analyses for my Lake Malawi population genomics study. Specifically, the alignments enable me to:

1. Distinguish ancestral vs. derived alleles at variant (segregating) sites in the Lake Malawi and Crater lake call sets
2. Assess long term evolutionary conservation of genomic regions of interest identified from the crater lake and Lake Malawi data
3. Use the Lake Victoria *P. nyererei* as a clear outgroup to root the phylogenetic tree of species within the Lake Malawi catchment
4. Find homologous sequences in zebrafish of regions of interest identified in the *M. zebra* genome, for example for follow-up functional studies

Producing (multiple) whole-genome alignments requires computational resources and expertise that are not available to a typical research group and was enabled by the strong computational facilities available at the Sanger Institute. Therefore, the multiple alignments in *P. nyererei*, *A. burtoni*, and *N. brichardi* genomic coordinates can be a valuable resource to research groups focussed on Lake Victoria and Lake Tanganyika cichlid radiations. Furthermore, the alignments will facilitate translation between genomic coordinates and thus enable comparisons between our findings based on alignment to the *M. zebra* genome, results produced by Lake Victoria researchers who use *P. nyererei* as the reference, and Lake Tanganyika results based on alignment to *A. burtoni* or *N. brichardi.*

Methods

Pairwise alignments of genome assemblies listed in Tables table:CichlidGenomes and table:alignGenomes were generated using `lastz v1.02` [139], with the following parameters:

For cichlid-cichlid alignments:

```
B=2 C=0 E=150 H=0 K=4500 L=3000 M=254 O=600 Q=human_chimp.v2.q T=2 Y=15000
```

For cichlid-other teleost alignments:

```
B=2 C=0 E=30 H=0 K=3000 L=3000 M=50 O=400 T=1 Y=9400
```

This was followed by using Jim Kent's `axtChain` tool with `-minScore=5000` for cichlid-cichlid and `-minScore=3000` for cichlid-other teleost alignments. Additional tools with default parameters were then used following the UCSC whole-genome alignment paradigm (`http://genomewiki.ucsc.edu/index.php/Whole_genome_alignment_howto`) in order to obtain a contiguous pairwise alignment.

Multiple alignment were generated from pairwise alignments using the `multiz v11.2` [140] program using default parameters and the following pre-determined phylogenetic tree: (((((((*M. zebra*, *P. nyererei*), *A. burtoni*), *N. brichardi*), *O. niloticus*), medaka), sticleback), zebrafish), in agreement with [93].

To obtain ancestral allele information for single nucleotide variants called against the *M. zebra* genome, indels were removed from the *M. zebra-P. nyererei* pairwise alignment (in *M. zebra* genomic coordinates), and ancestral allele information for variants filled into the `VCF` file using my custom C++ program `evo` (Available from `https://github.com/millanek/evo`) with the `aa-seq` and `aa-fill` options.

Table 3.4 **Versions of non-cichlid teleost assemblies used in whole-genome alignments.**

Species	UCSC version of assembly	URL used to download	Notes
medaka	oryLat2	ftp://hgdownload.soe.ucsc.edu/goldenPath/oryLat2/bigZips/oryLat2.fa.gz	NIG v1.0 assembly
stickleback	gasAcu1	ftp://hgdownload.soe.ucsc.edu/gbdb/gasAcu1/gasAcu1.2bit	Broad Institute v1.0
zebrafish	danRer7	http://hgdownload.cse.ucsc.edu/gbdb/danRer7/danRer7.2bit	Sanger Zv9 assembly

3.5 Cichlid genome browser

Introduction

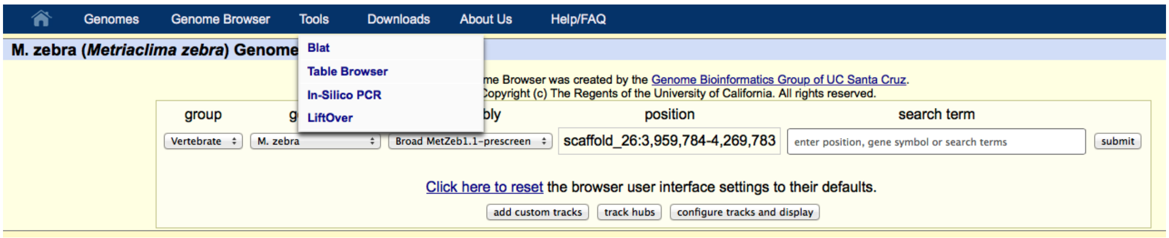
Interactive visual exploration is a key component of research with genomics datasets, complementing computational approaches [141]. Genome browsers enable users to examine a portion of the genome and various annotation tracks (e.g. assembly gaps, genes, repetitive sequence annotation, multiple sequence alignments) at arbitrary scale, from individual DNA bases up to a whole genome view. Several genome browsers with Web interfaces have been developed, originally hosting data and annotations related to the Human Genome Project [142]. Two of the most popular websites, Ensembl [17] (`http://www.ensembl.org`) and the UCSC Genome Browser [143] (`https://genome.ucsc.edu`) have since grown to host reference genomes and annotations for 77 and 69 vertebrate species respectively. However, of the reference genomes generated by the Cichlid Genome Consortium, only the *O. niloticus* assembly was deemed to be of sufficiently high quality for inclusion in the Ensembl and UCSC browsers. BouillaBase (`http://bouillabase.org`) at University of Maryland hosts genomes and annotations, and data from the Cichlid Genome Consortium, but its capabilities are limited compared to Ensembl or UCSC browsers, and at the time of writing it is very slow - to the point that I found it virtually unusable in support of my research.

An alternative to Web based global services has emerged in the form of standalone genome browsers enabling exploration of genomics datasets on standard desktop computers [141]. Rather than remotely presenting a pre-defined set of genomes and data, desktop browsers display datasets on users' computers and thus are much more flexible. The Integrative Genomics Viewer (IGV) [144] is perhaps the most popular of these tools and I have used it a number of times during the PhD. However, even these tools have significant drawbacks, including limitations on the amount of data that can be loaded (visualised genomes and data sets generally must be to be loaded in RAM memory), the need to re-load all datasets every time the program is restarted, and the inability to share data with collaborators.

To overcome the above difficulties and enable high quality visualisation for all cichlid genomes and data generated during this PhD, I set up the Cambridge Cichlid Browser (CCB). The CCB site runs on the UCSC Genome Browser engine, offers the majority of its functions, and is currently hosted on a server computer at the Gurdon Institute in Cambridge: `http://cichlid.gurdon.cam.ac.uk`.

Datasets and functions

Cambridge Cichlid Browser (CCB) hosts reference genomes and annotations generated by the Cichlid Genome Consortium and multiple datasets generated during my PhD. The browser is driven by an underlying MySQL database and the total volume of data available at the moment is ~100GB. CCB offers the majority of function of the UCSC browser, reviewed in [145]. In addition to exploring the five genomes and their annotations with zoom and scroll functions, it is possible to search by specifying genomic coordinates, search for genes by name, and search for homologous regions to a DNA or protein sequence with BLAT [146]. Other useful functions include 'Table Browser' for access to the underlying database, 'In-silico PCR' for fast design of PCR primers, 'LiftOver' for quick translation of genomic coordinates between reference genomes (based on whole-genome alignments), and PDF output of browser graphics. Figure 3.5 displays a screenshot from the browser's *M. zebra* genome gateway.



M. zebra (*Metriaclima zebra*) Genome

me Browser was created by the Genome Bioinformatics Group of UC Santa Cruz.
Copyright (c) The Regents of the University of California. All rights reserved.

group	g	bly	position	search term
Vertebrate	M. zebra	Broad MetZeb1.1-prescreen	scaffold_26:3,959,784-4,269,783	enter position, gene symbol or search terms

Click here to reset the browser user interface settings to their defaults.

Fig. 3.5 **Cambridge Cichlid Browser *M. zebra* genome gateway.**

Figure 3.6 shows an example of CCB graphics, showing ~900kb section of the 'scaffold 26' fragment of the *M. zebra* genome, with annotation tracks including assembly gaps, genes, and multiple alignment with cichlids and other teleosts.

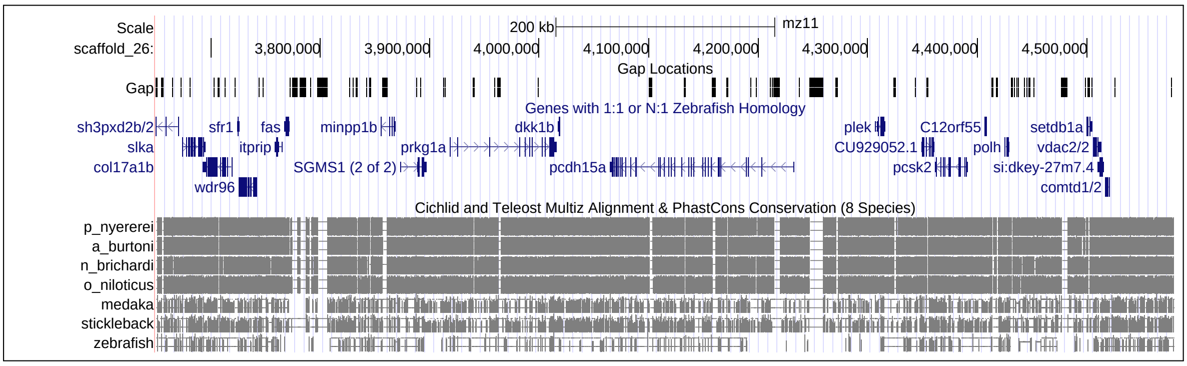


Fig. 3.6 **Cambridge Cichlid Browser - example browser graphic.**

Worldwide usage

The browser has been online since 1st March 2014 and has already made an impact beyond my immediate collaborators. Over 3,000 visits to the site have been recorded by the Google Analytics code I have linked to the website (`www.google.com/analytics/`). Limiting the statistics to session where users truly interacted with the browser (i.e. visited at least two pages, as opposed to just viewing the front page), there have been 657 browser sessions with a total of 19,607 page views initiated by 113 unique users (unique IP addresses). The average session duration has been 20 minutes and 37 seconds with an average 29.84 pages viewed per session.

The majority of the sessions have been initiated from the United States, UK, Germany, and Switzerland, but a number of other countries are also represented. Figures 3.7 and 3.8 show user locations and provide an insight into the user base. The locations correspond to major cichlid laboratories with a focus on genomics: for example, Craig Albertson's laboratory is currently located in Amherst, Massachusetts; Jeffrey Streelman's laboratory in Atlanta, Georgia; Russel Fernald's laboratory in Stanford; George Turner's laboratory in Bangor, Wales; and Axel Meyer's laboratory in Konstanz, Germany. In conclusion, these statistics make it clear that the Cambridge Cichlid Browser has already proven to be a valuable resource for the cichlid genomics research community.

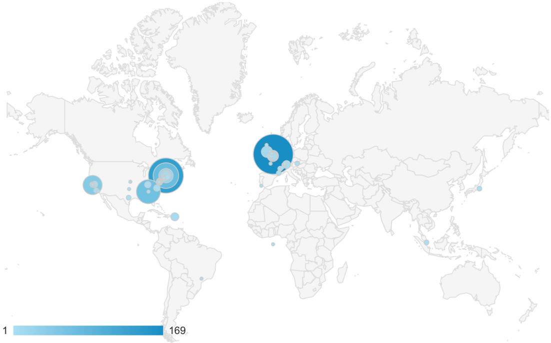
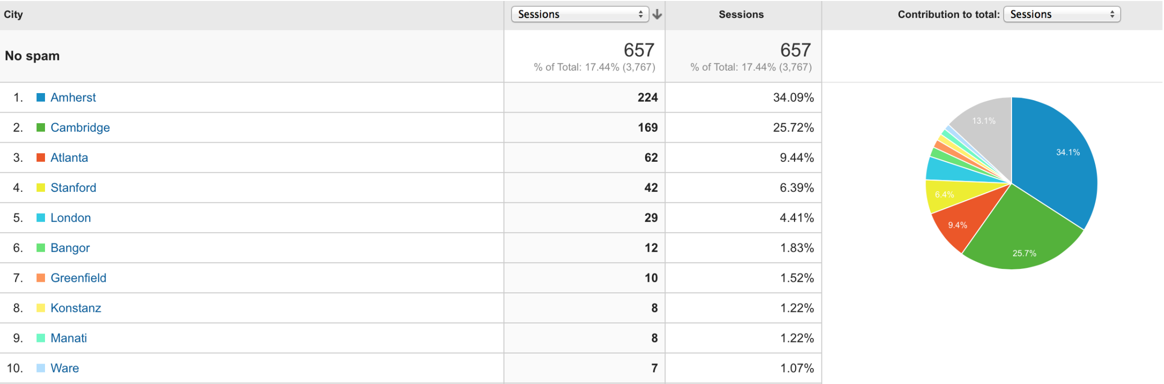


Fig. 3.7 **A map showing the location of CCB users.**



City	Sessions ↕ ↓	Sessions	Contribution to total: Sessions ↕
No spam	657 % of Total: 17.44% (3,767)	657 % of Total: 17.44% (3,767)	
1. 🟦 Amherst	**224**	34.09%	
2. 🟩 Cambridge	**169**	25.72%	
3. 🟥 Atlanta	**62**	9.44%	
4. 🟨 Stanford	**42**	6.39%	
5. 🟦 London	**29**	4.41%	
6. 🟩 Bangor	**12**	1.83%	
7. 🟧 Greenfield	**10**	1.52%	
8. 🟨 Konstanz	**8**	1.22%	
9. 🟩 Manati	**8**	1.22%	
10. 🟦 Ware	**7**	1.07%	

Fig. 3.8 **Ten cities with the highest contribution to CCB sessions.**