Supporting Information

Detailed Materials and Methods

Library Preparation, Gene Capture and Multiplexed Sequencing

Using a Covaris M220 sonicator (Gene Company Limited, Covaris, Woburn, MA), 300 ng of DNA from each sample was sheared to about 500 bp according to the manufacturer's instructions. DNA libraries were prepared following Meyer and Kircher (2010) with some modifications (Li et al. 2013). The amplified libraries were hybridized to the RNA baits following a cross-species gene enrichment protocol as recommended by Li et al. (2013). A "touch down" hybridization temperature scheme (starting from 65° C to 50° C reduced 5° C every 10 h) was applied. Each sample was captured twice under the same conditions to improve the enrichment results (Li et al. 2013). Custom designed 8 bp indices in the Illumina P7 adapter were utilized for multiplex sequencing. All samples, including the first and the second captured products from the same library, were pooled in equimolar ratio and a 10 pM sample mixture was loaded on one lane of an Illumina HiSeq 2000 run (2 × 108 bp paired-end reads, and 8 bp index read). The sequencing was performed by Novogene (Beijing, China).

Data Processing

Raw sequence reads from the Illumina platform were converted to fastq format and demultiplexed according to their indices using *bcl2fastq v1.8.3* (Illumina, San Diego, CA, USA). Sequences without index information were discarded. Sequences were trimmed using *Cutadapt* (Martin 2011) with *Trim_galore v0.2.8* (*http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/*) as a wrapper to remove the adapter sequences and sequences with a mean Phred quality score below 20, after which reads of the first capture and the second capture were merged. A custom Perl script was used to remove the duplicates resulting from PCR amplification and to parse the reads to each locus based on their similarity to the bait sequences. A *de novo*

assembly strategy implementing a two-step method was applied to assemble the filtered reads for each locus of each taxon, which combined a *de bruijn* graph approach with an overlap and consensus method. First, Trinity v20140717 (Grabherr et al. 2011) was used to complete a preliminary assembly. The *Trinity* runs were set as "--run_as_paired" and "--min_contig_length 100". Second, *Geneious v7.1.5* (http://www.geneious.com, Kearse et al., 2012) was used to further merge the loci containing more than one contig after *Trinity* assembling. A *Geneious* workflow with a minimum overlap of 15 bp and a minimum overlap identity of 95% was created for merging the contigs. Both of the merged contigs and contigs that could not be further merged using *Geneious* were pooled in the same file for subsequent analyses.

Salichos and Rokas (2011) compared different approaches and found that reciprocal blast is one of the best ways to identify orthologs, so we used an approximate strategy to retrieve orthologous sequences from the assembled contigs. First, each contig was aligned to its corresponding bait sequence using the Smith-Waterman algorithm (Smith and Waterman 1981). The contig with the highest score was chosen as the putative orthologous gene. The scoring matrix was set as follows: match = 1, mismatch = -1, gap = -1. Second, *blast v2.2.27* (Camacho et al. 2009) was used to align the recovered putative orthologous sequences against the genome of *O. niloticus* (the closest relative with a well-assembled genome, which was also used in bait design) to identify the potential paralogous sequences. If the best hit for a contig from a species was not in the target region in the genome of *O. niloticus*, it was excluded from further analyses, because the contig may be paralogous to the target sequence. A custom Perl script was used to batch process blast with the parameters: "-word_size 7 -gapopen 5 -gapextend 2 -penalty -1 -reward 1 -evalue 0.000001 -outfmt 6".

Finally, nucleotide alignments for each locus were performed using *Clustal Omega v1.1.1* (Sievers et al. 2011) with default parameter settings. Average pairwise distance was calculated for each locus using a custom Perl script. Genes with extraordinarily large p-dist values were checked by eye for mistakes in alignment, and then were manually corrected. The workflow of data processing is illustrated in Supplementary Fig S1.

Reads Mapping and Calculating Statistics for Gene Capture Results

To investigate the capture efficacy, the number of on-target reads and sequencing depth were calculated for each individual sample. Because none of the sampled species have a reference genome, the assembled homologous contig of each species were used as their reference. *BWA v0.7.5a* (Li and Durbin 2009) was used to map reads to the assemblies for each individual with the default parameters settings. *Samtools v0.1.19* (Li et al. 2009) was used to transform the output sam files to bam format and to retrieve statistics for gene mapping results.

References

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC bioinformatics 10:421.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29:644-652.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647-1649.

Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJP. 2013. Capturing protein-coding genes across highly divergent species. Biotechniques 54:321-326.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 17:10-12.

Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor protocols 2010:pdb prot5448.

Salichos L, Rokas A. 2011. Evaluating ortholog prediction algorithms in a yeast model clade. PLoS One 6:e18755.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J *et al.* 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 7:539.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol. 147:195-197.

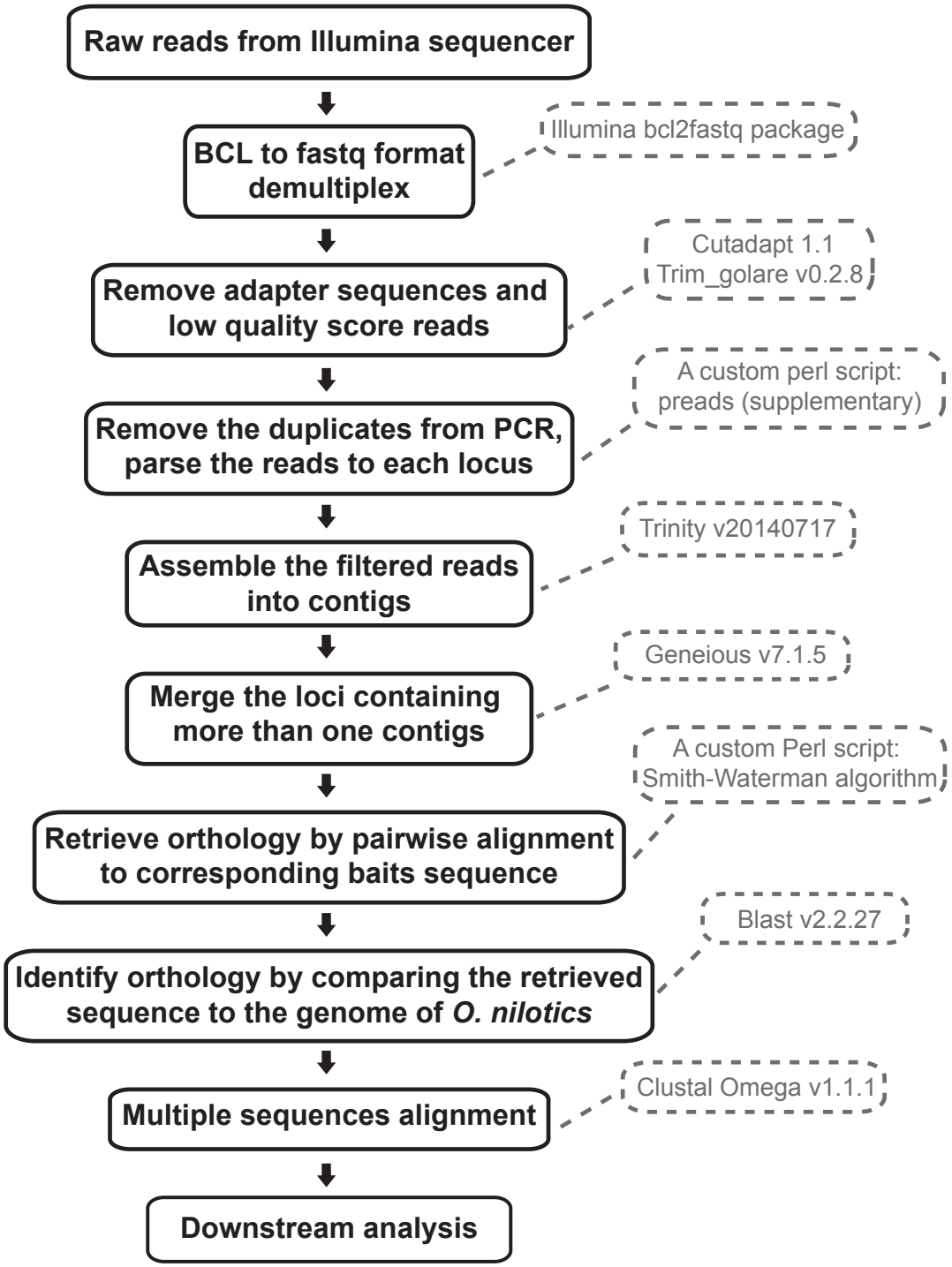

Fig. S1. Data assembly and alignment workflow.

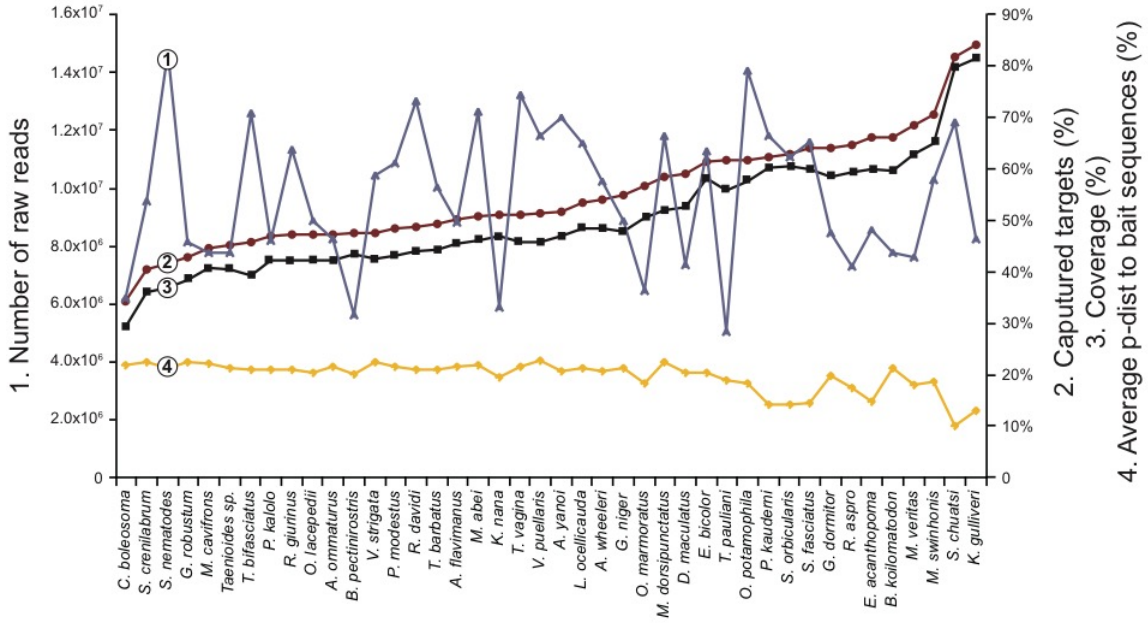

Fig. S2. Statistics related to efficacy of gene captures. (1) The number of raw reads (y-axis scale on the left side); (2) the percentage of captured targets (%); (3) proportion of the targets covered by reads (%); (4) average distance between the baits and the targets (%).

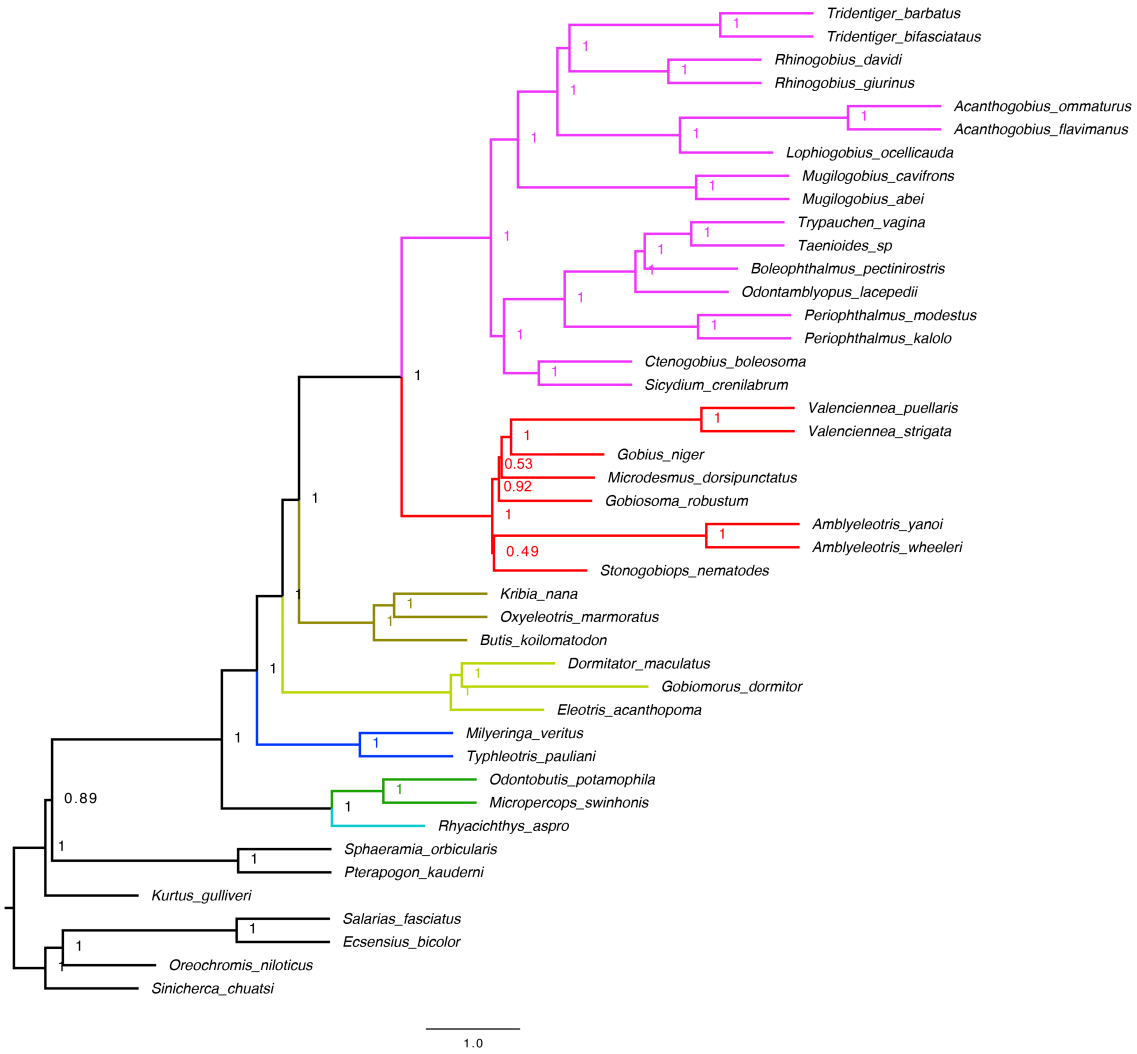

Fig. S3. ASTRAL species-tree from 570 loci with less than 5% missing data for all species. Support values are local posterior probabilities.

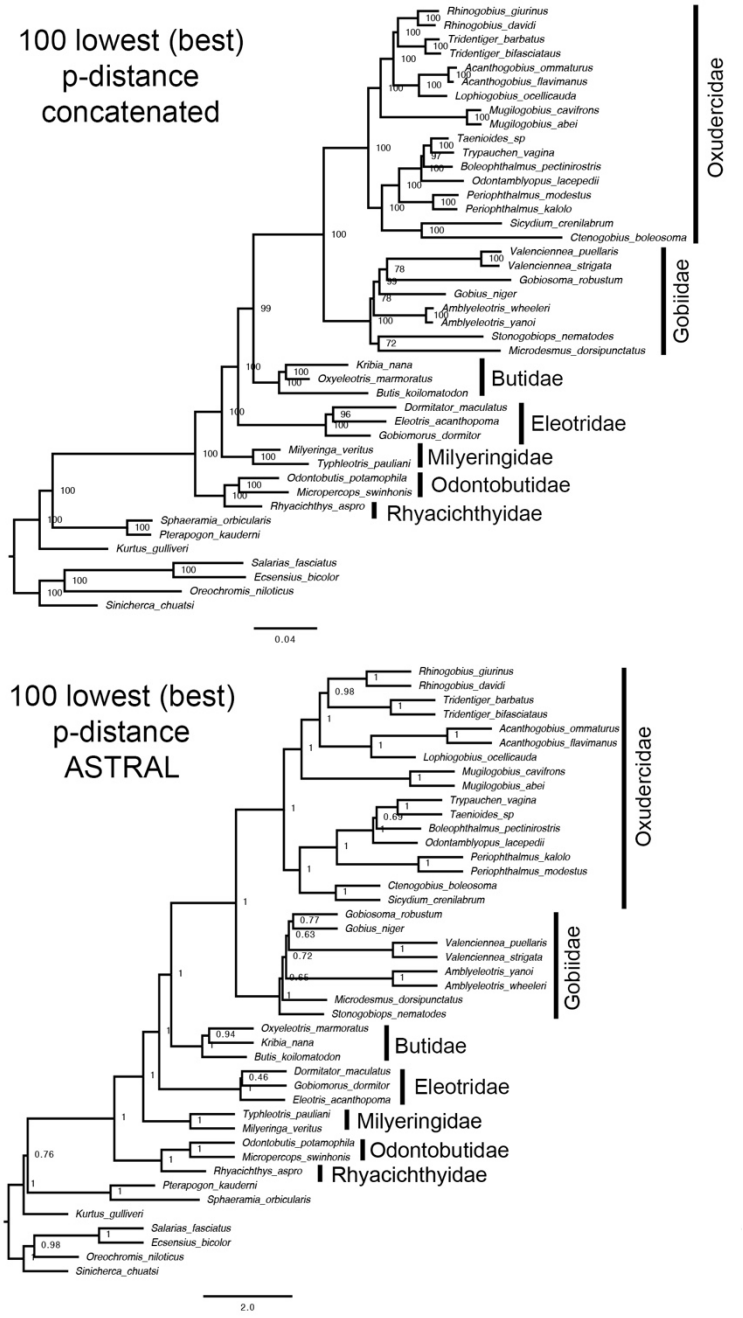

Fig. S4. Trees inferred from 100 loci with lowest average pairwise genetic distance (p-distance). (A) concatenated RAxML analysis, support at nodes are bootstrap values. (B) ASTRAL species-tree analysis, support at nodes are local posterior probabilities.

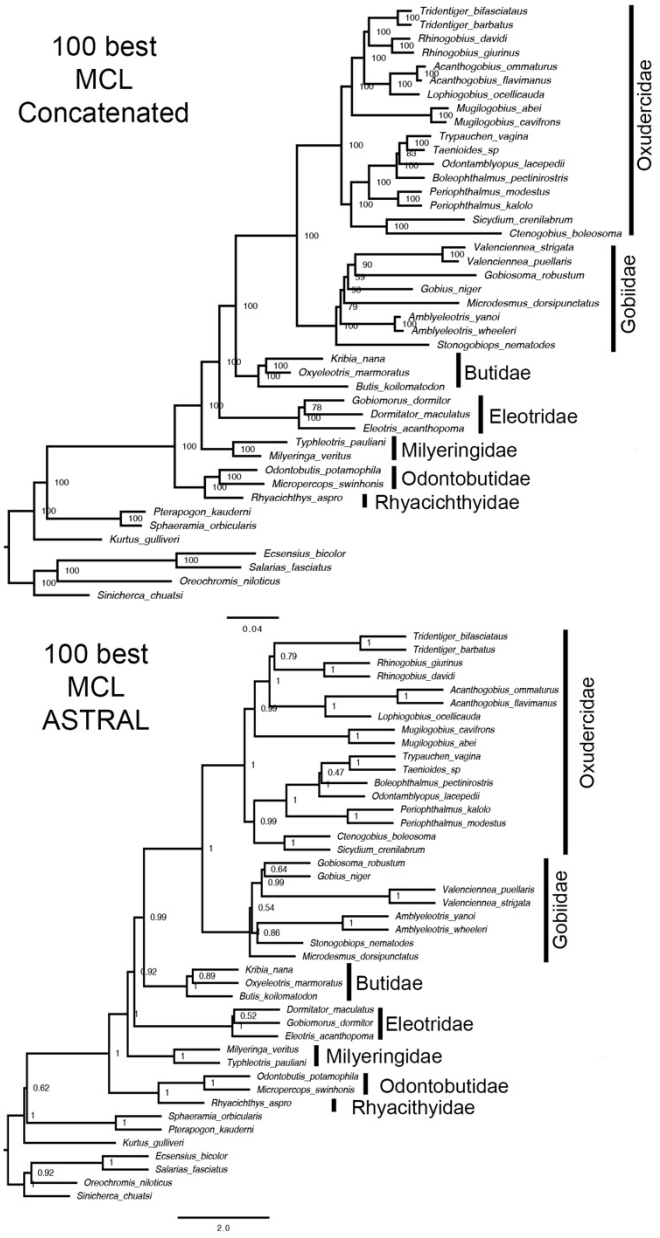

Fig. S5. Trees inferred from 100 most clocklike loci. (A) concatenated RAxML analysis, support at nodes are bootstrap values. (B) ASTRAL species-tree analysis, support at nodes are local posterior probabilities.

Table S1. Taxa sampled, summary statistics of the sequence reads and the assembly.

Taxon	No. of raw reads[1]	No. of filtered reads[2]	No. of assembled loci[3]		No. of loci after merging[4]		No. of captured targets (%)[5]	Total length	No. of mapped reads (%)	Coverage	Average Depth
			1 contig	≥ 2 contigs	1 contig	≥ 2 contigs					
1 Rhyacichthyidae											
Rhyacichthys aspro	7327646	7222068	4863	7534	5356	7041	11520 (64.7)	2086803	1505347 (20.8)	59.50%	57
2 Odontobutidae											
Odontobutis potamophila	14086352	13927990	6138	5374	6632	4880	11009 (61.8)	2024025	3546449 (25.5)	57.70%	138
Micropercops swinhonis	10309292	10170586	6332	6610	6905	6037	12557 (70.5)	2275694	1577689 (15.5)	64.90%	54
3 Milyeringidae											
Typhleotris pauliani	5056802	4924896	4512	7345	4936	6920	10982 (61.7)	1960487	368676 (7.5)	55.90%	15
Milyeringa veritas	7609098	7460506	4814	7817	5357	7271	12211 (68.5)	2208105	629225 (8.4)	62.90%	24
4 Eleotridae											
Gobiomorus dormitor	8468862	8365736	5806	6036	6293	5549	11441 (64.2)	2051703	1598602 (19.1)	58.50%	62
Eleotris acanthopoma	8574974	8344860	4548	7649	4991	7206	11780 (66.1)	2104720	576410 (6.9)	60.00%	21
Dormitator maculatus	7381820	7222658	4953	5949	5303	5599	10508 (59.0)	1850689	391982 (5.4)	52.80%	16
5 Butidae											
Butis koilomatodon	7762040	7582582	4680	7573	5107	7146	11797 (66.2)	2095152	550516 (7.3)	59.70%	20
Kribia nana	5895928	5815014	4286	5158	4673	4771	9085 (51.0)	1648589	1262000 (21.7)	47.00%	61
Oxyeleotris marmorata	6473574	6341852	4300	6585	4700	6185	10088 (56.6)	1772454	329258 (5.2)	50.50%	15

6 Gobiidae											
Amblyeleotris yanoi	12469922	12297616	4672	4915	5053	4534	9239 (51.9)	1651362	2515226 (20.5)	47.10%	118
Amblyeleotris wheeleri	10237774	10091056	4890	5083	5300	4673	9629 (54.0)	1702428	1769353 (17.5)	48.50%	81
Gobius niger	8867644	8763148	6186	4402	6610	3978	9803 (55.0)	1684048	378201 (4.3)	48.00%	16
Gobiosoma robustum	8129486	8015180	3801	4165	4127	3839	7642 (42.9)	1352208	1008878 (12.6)	38.50%	59
Valenciennea puellaris	11806568	11654278	5401	4577	5823	4155	9166 (51.5)	1611427	1960844 (16.8)	45.90%	95
Valenciennea strigata	10474966	10337514	4870	4302	5199	3972	8505 (47.7)	1493749	1622653 (15.7)	42.60%	83
Stonogobiops nematodes	14675334	14449244	3870	4139	4165	3844	7413 (41.6)	1301908	1630585 (11.3)	37.10%	90
“Microdesmidae”											
Microdesmus dorsipunctatus	11829892	11714268	6622	4185	7160	3647	10418 (58.5)	1826055	1987322 (17.0)	52.00%	83
7 Oxudercidae											
Acanthogobius flavimanus	8853528	8752236	5470	3862	5857	3475	8970 (50.4)	1600199	1478945 (16.9)	45.60%	71
Acanthogobius ommaturus	8263474	8160086	4767	4004	5150	3621	8429 (47.3)	1485623	1012768 (12.4)	42.40%	52
Ctenogobius boleosoma	6220112	6138248	2730	3663	2935	3458	6099 (34.2)	1023908	861092 (14.0)	29.20%	37
Mugilogobius cavifrons	7768992	7656136	3614	4681	3923	4372	7964 (44.7)	1430370	1146278 (15.0)	40.80%	63
Mugilogobius abei	12672064	12513750	4672	4736	4947	4461	9068 (50.9)	1618286	1599817 (12.8)	46.10%	76
Tridentiger barbatus	10037650	9907262	4701	4876	5131	4446	8783 (49.3)	1554832	1504006 (15.2)	44.30%	76
Tridentiger bifasciatus	12591712	12416362	3761	5092	4077	4776	8154 (45.8)	1382266	1727933 (13.9)	39.40%	92
Rhinogobius giurinus	11351640	11216416	5710	3253	5980	2983	8404 (47.2)	1480421	1448540 (12.9)	42.20%	70

Rhinogobius davidi	13040078	12851338	5100	4339	5522	3917	8691 (48. 8)	1544054	2414889 (18.8)	44.00%	122
Lophiogobius ocellicauda	11570950	11402942	5042	4811	5492	4361	9521 (53.4)	1702799	1952285 (17.1)	48.50%	89
Periophthalmus kalolo	8220646	8116210	4880	4212	5240	3852	8397 (47.1)	1483099	1021524 (12.6)	42.30%	53
Periophthalmus modestus	10860918	10723458	5574	3737	5933	3378	8618 (48.4)	1511654	1473615 (13.7)	43.10%	76
Boleophthalmus pectinirostris	5617656	9464262	4944	3914	5347	3511	8499 (47.7)	1526669	1322172 (14.0)	43.50%	67
Amblyopinae											
Taenioides sp.	7774864	7667718	3544	5218	3885	4877	8076 (45.3)	1416844	1108873 (14.5)	40.40%	62
Trypauchen vagina	13248270	13079186	5584	4240	6001	3823	9098 (51.0)	1606929	1912897 (14.6)	45.80%	88
Odontamblyopus lacepedii	8876558	8774886	4930	3840	5243	3527	8416 (47.2)	1482855	1014960 (11.6)	42.30%	51
Sicydiinae											
Sicydium crenilabrum	9587514	9464262	3416	4378	3691	4103	7217 (40.5)	1261145	940293 (9.9)	36.00%	75
8 Kurtidae											
Kurtus gulliveri	8225564	8007234	5236	10149	5982	9402	15003 (84.2)	2860749	950758 (11.9)	81.50%	28
9 Apogonidae											
Sphaeramia orbicularis	11081578	10918472	6096	6075	6729	5442	11205 (62.9)	2128345	2781642 (25.5)	60.70%	104
Pterapogon kauderni	11810784	11656832	6216	5802	6819	5199	11087 (62.2)	2113946	2734312 (23.5)	60.30%	97
10 Blenniidae											
Salarias fasciatus	11618660	11412778	5634	6788	6264	6158	11405 (64.0)	2103614	2824070 (24.7)	60.00%	102
Ecsenius bicolor	11299600	11123854	6259	6415	6742	4538	10929 (61.3)	2044323	2171807 (19.5)	58.30%	84

11 Sinipercidae											
Siniperca chuatsi	12291914	12114602	9502	6257	10380	5379	14567 (81.8)	2793510	3612188 (29.8)	79.60%	106
12 Cichlidae											
Oreochromis niloticus[6]	--	--	--	--	--	--	17817 (100)	3508171	--	--	--
Average	9769588	9719990	5070	5327	5499	4864	9795 (55.0)	1758524	1529163 (15.1)	50.10%	67

[1] number of raw reads; [2] number of reads after filtered out reads with low quality score; [3] number of loci with one contig and more than two contigs respectively after *de novo* assembly using Trinity; [4] number of loci with one contig and more than two contigs respectively, after further merging using Geneious; [5] number of captured orthologous CDS loci and the percentage of captured CDS to total targets; [6] sequences download from Ensembl database.

Table S2. Correlations between values (raw) for each of the two predictors and the length for 570 genes with less than 5% missing data. Bold=significant at $p < 0.01$.

Variables	p	R^2
MCL	**2.2e-16**	**0.6923**
p-distance	**1.818e-8**	**0.0674**

Table S3. Comparison of alternative multiple regression models

Alternative models	Predictors included	Significant predictors (p<.005)	Predictors with significant interaction terms with locus length (p<.005)	Adjusted R squared	p-value
Full Model, after correcting for locus length	p distance (length corrected), MCL (length corrected)	p distance (length corrected), MCL (length corrected),	N/A	0.1516	2.20E-16
Full model, no corrections, including length interaction	p distance, MCL, locus length	MCL, locus length	MCL, p-distance	0.1707	2.20E-16
Full model, no corrections, log transformed predictors, including length interaction	log(p distance), log(MCL), locus length	log(MCL), locus length	log(MCL), log(p distance)	0.1986	2.20E-16

Supplementary data file 1.
Custom Perl scripts (Supplementary_data_file_1.txt).

Supplementary data file 2.
Raw values for each characteristic for each gene (Supplementary_data_file_2.txt).

Supplementary data file 3.
Sequence data for the 570 loci (Supplementary_data_file_3.txt)