

# Rapid isolation of CA microsatellites from the tilapia genome

K. L. Carleton, J. T. Streebman, B.-Y. Lee, N. Garnhart, M. Kidd and T. D. Kocher

Hubbard Center for Genome Studies, University of New Hampshire, Durham, NH, UK

---

## Abstract

We have developed  $(CA)_n$  microsatellite markers for the cichlid fish, *Oreochromis niloticus* using a variation of the hybrid capture method. The resulting genomic library was highly enriched in repetitive DNA with 96% of clones containing CA repeats. The number of repeats ranged from four to 45 with an average of 19. Two-thirds of the sequenced clones had 12 or more repeats and sufficient flanking sequence to design primers. The resulting markers were tested in an  $F_2$  cross of *O. niloticus*  $\times$  *O. aureus*. Nearly 90% of the markers amplified in this cross and 74% of these were informative. This work demonstrates the importance of minimizing the number of polymerase chain reaction (PCR) amplification cycles before and after the enrichment steps to reduce PCR recombination and the generation of chimaeric clones.

**Keywords** CA repeat, linkage mapping, marker generation, microsatellite.

---

Numerous methods have been developed to enrich genomic libraries for particular microsatellite classes. The earlier primer extension methods (e.g. Ostrander *et al.* 1992) are technically demanding. More recently, simpler hybrid capture methods have been developed (e.g. Kandpal *et al.* 1994). These methods use a biotinylated  $(CA)_n$  probe to capture repetitive sequences and immobilize them on streptavidin-coated magnetic beads. After several wash steps, the captured DNA is eluted, amplified and cloned to produce a library enriched for the target sequence. In this work, we build on the methods of Prochazka (1996), Brown *et al.* (1995) and Moraga Amador *et al.* (1998) to isolate  $(CA)_n$  microsatellite markers for linkage mapping in tilapia (Kocher *et al.* 1998; Lee *et al.* unpublished observation).

Several enriched libraries were constructed. These libraries differed in the selected size of the restricted genomic DNA, the linker used, and how much they were polymerase chain reaction (PCR) amplified prior to the probe hybridization step (Table 1). Genomic DNA was isolated from muscle tissue of an individual *O. niloticus* using a standard proteinase K/phenol extraction protocol. The DNA (5–10  $\mu$ g) was then treated with RNase and restricted with *Dpn*II. Size fractions of either 300–600 bp or 400–800 bp

were recovered from a 1% agarose gel using a Qiaquick gel extraction column (Qiagen Inc., Valencia, CA, USA). Linkers were ligated onto the DNA fragments using T4 DNA ligase. Excess linker was removed by washing on an Ultrafree column (Millipore, Bedford, MA, USA). Pre-hybridization PCR amplification was then performed for libraries 1, 3, and 4 using one of the linker oligos (see Table 1).

For enrichment, the DNA was denatured and hybridized to a biotinylated probe [B-ATAGAATAT(CA)<sub>16</sub>] in 3 $\times$  SSC/0.1% SDS at 68 °C for 1 h. The DNA hybridized to the probe was then captured with streptavidin magnetic beads and washed (twice in 6 $\times$  SSC/0.1% SDS at room temperature, twice in 3 $\times$  SSC/0.1% SDS at 68 °C, and twice in 6 $\times$  SSC at room temperature). Following purification, the DNA was denatured from the beads in 0.1 $\times$  TE at 95 °C and PCR amplified including a final 10 min extension time. The DNA was then spin-cleaned on Ultrafree columns (Millipore), quantified and TA cloned into pGEM-T (Promega, Madison, WI, USA). The ligation was used to transform JM109 cells and plated.

Each of the libraries was picked into 60  $\mu$ l of LB/AMP with 7.5% glycerol in 96 well plates and grown for several hours at 37 °C. Clones were PCR tested for insert size. Those with inserts >250 bp were used to inoculate 2–3 ml overnight cultures and minipreped to extract plasmid DNA (Qiagen). Each clone was sequenced in one direction (ABI 377). After sequence analysis, a small fraction of the clones required sequencing of the opposite strand.

Address for correspondence

Karen L. Carleton, Hubbard Center for Genome Studies, 35 Colovos Road, University of New Hampshire, Durham NH 03824, UK.  
E-mail: karen.carleton@unh.edu

Accepted 9 October 2001

**Table 1** Libraries enriched for microsatellite DNA.

Library	1	2	3	4
Restriction size	300–800	300–600	300–600	400–800
Linker	CD <sup>1</sup>	AB <sup>2</sup>	AB <sup>2</sup>	AB <sup>2</sup>
Pre-hyb PCR				
Cycles	30	0	10	10
Anneal Tm	54	–	68	68
Post-hyb PCR				
Cycles	30	15	15	15
Anneal Tm	54	68	68	68

<sup>1</sup>Linker CD anneals together oligo C: 5'GTCAAGAATTCGGTACCGT-CGAC and oligo D 5'GATCGTCGACGGTACCGAATTCT (Brown *et al.* 1995). PCR uses oligo C.

<sup>2</sup>Linker AB anneals together oligo A: 5'GCGGTACCCGGGAAGCTTGG and 5' phosphorylated oligo B: 5'GATCCCAAGCTTCCCGGTACCGC (Moraga Amador *et al.* 1998). PCR uses oligo A.

The resulting sequences were examined for the presence of microsatellites. Sequences were selected which contained perfect repeats at least 12 repeats in length, or interrupted or compound repeats with a total of at least 12 repeats and no more than one non-repetitive pair in the repeat [e.g. (CA)<sub>8</sub>AA(CA)<sub>4</sub>]. Repeat length for each locus was recorded as the longest stretch of uninterrupted CA's. Data for each clone were compiled in a Filemaker database to facilitate batch processing via BLAST (Altschul *et al.* 1990) and automated submission to GenBank. Duplicate clones and those which matched known minisatellite sequences were not analysed further.

Primers were designed for sequences meeting the selection criteria (Primer 3, Rozen & Skaletsky 1998). All primers were designed for 60 °C annealing temperature and a total amplicon size between 100 and 250 bp to simplify multiplex PCR and rapid gel separation. One primer in each pair was fluorescently labelled with either Tet, 6-Fam, or Hex (Operon Technologies Inc., Alameda, CA, USA). Each primer pair was tested on the genomic DNA used to generate the library as well as an F2 cross between *Oreochromis niloticus* × *O. aureus* using standard conditions (94 °C 20 s, 50–55 °C 30 s, 72 °C 1 min, 30 cycles). Fragments were resolved on 4% gels using an ABI 377 automated DNA sequencer.

Of the 373 clones sequenced from library 1, 64% (238 clones) contained simple or compound microsatellites with 12 or more repeats. Primers were designed for 35 of these loci and tested on genomic DNA. Only seven (20%) successfully amplified genomic DNA despite the fact that all of the primer pairs successfully amplified the plasmid DNA of the clone from which they were designed. Five of the seven were informative in the *Oreochromis* cross and are deposited in GenBank (G63968, G63984, G63987, G64035 and G64061).

The low success rate of PCR primers derived from library 1 suggested that the majority of clones were chimaeric and not representative of contiguous stretches of genomic DNA. The DNA for library 1 was prepared with two PCR amplification steps, each of 30 cycles. The repetitive nature of microsatellite DNA could have led to recombinant PCR at each of these steps (Bradley & Hillis 1997).

We, therefore, greatly reduced the number of PCR cycles used to produce subsequent libraries. The pre-hybridization PCR amplification selects for DNA fragments which have linkers attached to both ends, although the majority of template DNA should have both linkers. Library 2 was, therefore, generated without any amplification at this step and only 10 cycles were used to amplify libraries 3 and 4. Post-hybridization PCR amplification generates sufficient double stranded DNA with a terminal A nucleotide for TA cloning. We determined that 15 cycles of amplification were sufficient for this purpose. Both the pre- and post-hybridization amplifications were performed using a mixture of *Taq* polymerase and a proof reading enzyme (Dynazyme EXT, MJ Research) to reduce extension from mismatched templates.

Comparison of libraries 2, 3 and 4 showed that there was little difference in the success rate for converting clones to informative markers. Libraries 2 and 3 were generated from similarly sized fractions (300–600 bp) and their results were combined. For the 99 clones sequenced from these libraries, the average insert size was 390 ± 115 bp. Of these, 19 clones (19%) had too little sequence flanking the microsatellite to design primers. For library 4, the initial genomic DNA was sized at 400–800 bp. For the 141 clones sequenced from this library, the average insert size was 490 ± 165 bp. Of these, 18 clones (13%) had insufficient flanking sequence to design primers. Therefore, selecting slightly larger genomic DNA fragments did increase the probability that sequenced clones could be converted to microsatellite markers.

Combining the results from these three libraries, 35% of the sequenced clones did not meet the selection criteria for primer design. This group includes clones which had no repeat (3.9%), clones with insufficient flanking sequence (14.5%), clones whose repeats were too short (9.8%), duplicate clones (9.8%), and clones whose repeats were too complex or that matched known minisatellite sequences (3.9% which matched GenBank #AF016497, Takahashi *et al.* 1998; or #AB009705). Primers were designed for the remaining 65% (165 clones) for testing. Of these 88% (145 clones) successfully amplified genomic DNA. This suggested that using fewer cycles of PCR amplifications successfully reduced the number of recombinant clones.

Of the marker pairs which amplified cleanly, 108 (74%) were informative in an interspecific F<sub>2</sub> reference family (*O. niloticus* × *O. aureus*). Of the remaining 37 markers

**Table 2** Polymerase chain reaction (PCR) amplification conditions and allele sizes for *Oreochromis* markers. Markers are identified by GenBank accession number and sequence tagged site (STS) number. Allele sizes for the *Oreochromis*  $F_2$  family are listed as is the PCR annealing temperature.

GenBank	STS#	Allele sizes	PCR temperature °C	GenBank	STS#	Allele sizes	PCR temperature °C
G63968	UNH719	124, 126	50	G68237	UNH931	199, 203, 240	53
G63984	UNH735	168, 178, 182	50	G68238	UNH932	122, 124, 143	53
G63987	UNH738	157, 171	50	G68239	UNH933	241, 243, 245	53
G64035	UNH817	99, 119	50	G68240	UNH934	221, 241	53
G64061	UNH773	207, 211, 253	50	G68241	UNH937	187, 193	53
G68180	UNH840	131, 138, 140	55	G68242	UNH940	153, 157, 189	53
G68181	UNH841	114, 156	52	G68243	UNH942	133, 135	53
G68182	UNH843	109, 120, 124	55	G68244	UNH943	137, 146	53
G68183	UNH844	98, 104, 116	55	G68245	UNH946	153, 178	55
G68184	UNH845	178, 181	55	G68246	UNH948	180, 198	55
G68185	UNH846	173, 203	55	G68247	UNH949	154, 158	55
G68186	UNH848	180, 192, 205	55	G68248	UNH951	195, 206	55
G68187	UNH849	155, 191	52	G68249	UNH952	191, 193	55
G68188	UNH851	111, 117, 122	55	G68250	UNH954	135, 185	55
G68189	UNH853	172, 174, 186	55	G68251	UNH957	161, 174, 176	55
G68190	UNH854	225, 229	52	G68252	UNH958	143, 153, 155	55
G68191	UNH855	152, 162	55	G68253	UNH960	130, 177	55
G68192	UNH856	189, 197	55	G68254	UNH961	191, 195	55
G68193	UNH857	147, 179	52	G68255	UNH965	174, 179, 192	55
G68194	UNH858	257, 280	50	G68256	UNH967	116, 161, 217	55
G68195	UNH860	191, 229, 233, 241	55	G68257	UNH968	196, 201, 209	55
G68196	UNH863	140, 164, 169	55	G68258	UNH970	89, 94, 114, 152	55
G68197	UNH865	218, 228	50	G68259	UNH971	201, 214, 234	55
G68198	UNH866	153, 162	55	G68260	UNH973	127, 132, 136, 147	55
G68199	UNH868	216, 220	50	G68261	UNH974	177, 185, 222	55
G68200	UNH869	149, 151	55	G68262	UNH977	132, 143, 152	55
G68201	UNH871	219, 328	50	G68263	UNH979	239, 251, 269	55
G68202	UNH874	209, 219, 225, 236	55	G68264	UNH980	212, 221, 223	55
G68203	UNH875	135, 150	52	G68265	UNH982	106, 120, 130	55
G68204	UNH876	217, 247, 249	52	G68266	UNH985	133, 155	55
G68205	UNH878	155, 166, 171	55	G68267	UNH986	193, 204, 266	55
G68206	UNH879	195, 203	55	G68268	UNH988	204, 207, 210	55
G68207	UNH880	146, 182, 204	55	G68269	UNH989	142, 150, 152	55
G68208	UNH884	105, 127, 145	55	G68270	UNH990	150, 155, 162	55
G68209	UNH886	163, 172, 174	55	G68271	UNH991	164, 166	55
G68210	UNH887	155, 161, 172, 178	55	G68272	UNH993	161, 186, 190	55
G68211	UNH888	212, 319	55	G68273	UNH994	204, 215, 223	55
G68212	UNH890	244, 254, 270	52	G68274	UNH995	219, 236	55
G68213	UNH891	159, 167, 172	55	G68275	UNH996	194, 204, 206	55
G68214	UNH896	157, 227, 234	55	G68276	UNH997	125, 130	55
G68215	UNH898	274, 280	55	G68277	UNH998	112, 116, 120	55
G68216	UNH899	146, 166, 172	55	G68278	UNH999	105, 108, 110	55
G68217	UNH901	145, 159, 161	55	G68279	UNH1000	136, 143	55
G68218	UNH904	134, 136, 146, 164	55	G68280	UNH1003	160, 167, 189	55
G68219	UNH905	146, 151, 198	55	G68281	UNH1004	166, 177, 185	55
G68220	UNH906	152, 154, 156	55	G68282	UNH1005	144, 157	55
G68221	UNH907	122, 139	55	G68283	UNH1007	155, 247	55
G68222	UNH908	107, 163, 178	55	G68284	UNH1008	95, 99	55
G68223	UNH910	93, 107	55	G68285	UNH1009	148, 179, 189	55
G68224	UNH911	140, 144, 147	55	G68323	UNH918	105, 134	55
G68225	UNH913	98, 109, 115	55	G68324	UNH909	238, 246, 286	50

Table 2 (Continued).

GenBank	STS#	Allele sizes	PCR temperature °C	GenBank	STS#	Allele sizes	PCR temperature °C
G68226	UNH914	146, 162, 183	50	G68287	UNH873	203	50
G68227	UNH915	145, 147, 153	55	G68288	UNH883	224	55
G68228	UNH916	135, 142, 163	55	G68289	UNH889	191	55
G68229	UNH917	197, 222	55	G68290	UNH892	148	50
G68230	UNH919	182, 188, 192	53	G68291	UNH900	203	52
G68231	UNH920	150, 260	50	G68292	UNH924	223	48
G68232	UNH921	199, 211, 215	53	G68293	UNH935	185	50
G68233	UNH923	131, 151	55	G68294	UNH964	124	55
G68234	UNH925	220, 242	55	G68295	UNH978	171	55
G68235	UNH927	196, 204, 229	53	G68296	UNH987	167	55
G68236	UNH929	134, 172	53	G68297	UNH992	130	55

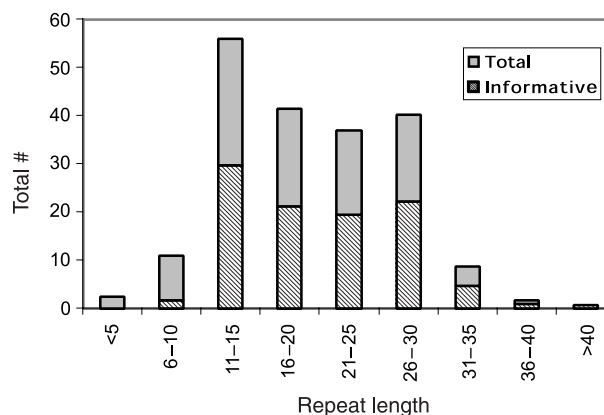
(26%), 11 markers were homozygous. The other 26 markers were difficult to score because of multiple, complex peaks, although many of these were also invariant in the cross. The informative markers (G68180-G68285, G68323-4), those that were homozygous (G68287-97), and those that were complex (G68286 and G68298-G68322) have been deposited in the GenBank STS database. The allele sizes in the  $F_2$  *Oreochromis* family and the PCR annealing temperatures for the informative and homozygous markers are listed in Table 2.

Of the 145 loci which were genotyped in the  $F_2$  cross, nine had Blastn hits with expectations less than  $10^{-5}$ ; four of these had expectations less than  $10^{-10}$ . Of these four, two of the markers were informative and will be placed on the tilapia linkage map (Lee *et al.* unpublished observation). This includes G68249/UNH952 which blasted to XM\_048114 ( $E = 2 \times 10^{-12}$ ) in humans and encodes a cytokine like nuclear factor (nPAC) and G68254/UNH961 which blasted to AF251021 ( $E = 2 \times 10^{-17}$ ) in mouse and encodes phosphofructokinase-1. This indicates it is possible to link a few of the markers with gene homologues through Blast, although with a low success rate.

The size distribution of all 238 clones which contain repeats are shown in Fig. 1. The average clone contained 19 repeat units. The size distribution is not necessarily indicative of the size of all genomic CA repeats because of the selection imposed by the hybridization procedure. The distribution is shifted to shorter sizes in comparison to cold water teleosts (Brooker *et al.* 1994).

The size distribution of the loci which are informative in the *Oreochromis* cross is also plotted in Fig. 1. These loci are evenly distributed across the range of repeat size. The average informative locus had 20 repeat units. This suggests that our criteria for selecting sequences (12 or more repeats) were sufficient to identify polymorphic loci. Shorter

CA microsatellite size distribution



**Figure 1** Distribution of microsatellite loci developed by the hybrid capture method. The loci are grouped by repeat length. Data is given for all loci generated by the method as well as those loci which are informative in the *Oreochromis* cross examined. The average repeat length is 19 while the average repeat length for informative loci is 20.

repeats are just as likely to be informative as longer ones, and are typically easier to score.

The libraries generated by this hybrid capture method were highly enriched for microsatellites; 96% of the clones contained  $(CA)_n$  repetitive sequence. Of the sequenced clones, 65% had repeats of sufficient length and flanking sequence to design primers. These repetitive sequences were broadly distributed with a mean repeat size of 19. If PCR amplifications are kept to 15 cycles or less during construction of the libraries, few chimaeric clones are generated. In our experience, 88% of primers designed successfully amplified genomic DNA. Of these, approximately 74% were polymorphic in the interspecific reference cross. Therefore, 42% of the clones that were sequenced

resulted in informative markers for mapping. These informative markers had a size distribution similar to the total set. The protocol described here can be completed in less than a month, provides an unbiased source of polymorphic microsatellite markers, and can be applied to nearly any organism for linkage mapping or population genetics.

### Acknowledgments

We thank Gideon Hulata for producing the  $F_2$  reference family. JTS was supported by an Alfred P. Sloan Foundation/NSF Postdoctoral Fellowship in Molecular Evolution, DBI 98-03946. This work was supported by grants from the USDA NRICGP (#98-03476) and the UNH Agricultural Experiment Station. This is HCGS contribution #1.

### References

- Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–10.
- Bradley R.D. & Hillis D.M. (1997) Recombinant DNA sequences generated by PCR amplification. *Molecular Biology Evolution* **14**, 592–3.
- Brooker A.L., Cook D., Bentzen P., Wright J.M. & Doyle R.W. (1994) Organization of microsatellites differs between mammals and cold-water teleost fishes. *Canadian Journal of Fisheries and Aquatic Science* **51**, 1959–65.
- Brown J., Hardwick L.J. & Wright A.F. (1995) A simple method for rapid isolation of microsatellites from yeast artificial chromosomes. *Molecular and Cellular Probes* **9**, 53–8.
- Kandpal R., Kandpal G. & Weissman S.M. (1994) Construction of libraries enriched for sequence repeats and jumping clones, and hybridization selection for region-specific markers. *Proceedings of the National Academy of Sciences USA* **91**, 88–92.
- Kocher T.D., Lee W.J., Sobolewska H., Penman D. & McAndrew B. (1998) A genetic linkage map of a cichlid fish, the tilapia (*Oreochromis niloticus*). *Genetics* **148**, 1225–32.
- Moraga Amador D., Farmerie B., Holland K., Blake A., Brazeau D. & Whitten M. (1998) *Tools for Developing Molecular Markers: Microsatellite Libraries and AFLPS. Workshop Manual*. University of Florida, Gainesville, FL.
- Ostrander E.A., Jong J.M., Rine J. & Duyk G. (1992) Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. *Proceedings of the National Academy of Sciences USA* **89**, 3419–23.
- Prochazka M. (1996) Microsatellite hybrid capture technique for simultaneous isolation of various STR markers. *Genome Research* **6**, 646–9.
- Rozen S. & Skaletsky H. J. (1998). Primer 3. Code available at: [http://www-genome.wi.mit.edu/genome\\_software/other/primer3.html](http://www-genome.wi.mit.edu/genome_software/other/primer3.html)
- Takahashi K., Terai Y., Nishida M. & Okada N. (1998) A novel family of short interspersed repetitive elements (SINEs) from cichlids: the patterns of insertion of SINEs at orthologous loci support the proposed monophyly of four major groups of cichlid fishes in Lake Tanganyika. *Molecular Biology Evolution* **15**, 391–407.