Mar Biotechnol
DOI 10.1007/s10126-009-9236-8

SHORT COMMUNICATION

Identification of Repetitive Elements in the Genome of *Oreochromis niloticus*: Tilapia Repeat Masker

Andrey Shirak · Manfred Grabherr · Federica Di Palma · Kerstin Lindblad-Toh · Gideon Hulata · Micha Ron · Tom D. Kocher · Eyal Seroussi

Received: 14 July 2009 / Accepted: 30 October 2009

Abstract The large-scale bacterial artificial chromosome-end sequencing project of Nile tilapia (*Oreochromis niloticus*) has generated extensive sequence data that allowed the examination of the repeat content in this fish genome and building of a repeat library specific for this species. This library was established based on Tilapiini repeat sequences from GenBank, sequences orthologous to the repeat library of zebrafish in Repbase, and novel repeats detected by genome analysis using MIRA assembler. We estimate that repeats constitute about 14% of the tilapia genome and also give estimates for the occurrence of the different repeats based on the Basic Local Alignment Search Tool searches within the database of known tilapia sequences. The frequent occurrence of novel repeats in the tilapia genome indicates the importance of using the species-specific repeat masker prior to sequence analyses. A web tool based on the RepeatMasker software was designed to assist tilapia genomics.

Keywords Transposable element · Sequence assembly · Repeat masking

A. Shirak · G. Hulata · M. Ron · E. Seroussi (✉)
Agricultural Research Organization, Institute of Animal Science,
Bet Dagan 50250, Israel
e-mail: Seroussi@agri.huji.ac.il

M. Grabherr · F. Di Palma · K. Lindblad-Toh
Broad Institute of Harvard and MIT,
320 Charles Street,
Cambridge, MA 02141, USA

T. D. Kocher
Department of Biology, University of Maryland,
College Park, MD 20742, USA

Introduction

Repeat masking is a crucial step in many sequence analyses including assembly of genomic and expressed sequence tag sequences (Tang 2007; Malde and Jonassen 2008), sequence searches as well as gene identification and annotation (Smith et al. 2007), and the design of PCR primers and hybridization probes (Andreson et al. 2006). However, repeat libraries are not available for most fish species, and it is a common practice to mask against known repeats from other model organisms such as zebrafish (*Danio rerio*) and pufferfish (*Takifugu rubripes*), which is less effective than masking with repeat libraries that are species-specific (Malde and Jonassen 2008). Several classes of repeats have been described in cichlid fish mostly in *Oreochromis niloticus*. They include satellite DNAs (Oliveira and Wright 1998), long interspersed nuclear elements (LINEs; Oliveira et al. 1999), telomeric $(TTAGGG)_n$ repeats (Chew et al. 2002), rDNA repeats (Martins et al. 2002), short interspersed repetitive elements (SINEs; Terai et al. 2003), and heterochromatic repetitive sequences (Ferreira and Martins 2008; Mazzuchelli and Martins 2009). This work has annotated a repeat library for tilapia by combining the previously annotated repeats from Tilapiini, sequences from *O. niloticus* bacterial artificial chromosome (BAC)-end project that were orthologous to the zebrafish repeat library, and novel repeats of tilapia classified by genome sequence analysis.

Materials and Methods

Preparation of Tilapia Repeat Library

GenBank records (209) of annotated repetitive sequence in Tilapiini were located (http://www.ncbi.nlm.nih.gov/) with-

Published online: 21 November 2009

in 231 non-mRNA sequences that corresponded to the following limits ((Tilapiini[Organism]) AND ((repeat[All Fields]) OR (transposon*[All Fields]) OR (repetitive[All Fields]) OR (SINE*[All Fields]) OR (LINE*[All Fields]) OR (satellite*[All Fields]) OR ("ribosomal RNA"[All Fields]))) NOT (mitochondri*[All Fields]). We frequently encountered failure to identify and remove all of the vector sequence in the finished BAC-end sequences (Fig. 1a). In order to detect such sequence contamination common to the current genomic projects of this fish tribe, the BAC cloning vector (FJ160466) was added to our repeat library records. BAC-end sequences (153,216) of *O. niloticus* were downloaded from Trace Archive (CENTER_PROJECT = "G1447", http://www.ncbi.nlm.nih.gov/Traces/) and combined with 7,855 Tilapiini records available from GenBank to form a local database that was searchable by Basic Local Alignment Search Tool (BLAST; blastall 2.2.17) on an Ubuntu (8.04 Hardy) 64-bit Q6600 Linux machine. *D. rerio* records for high complexity repeats in Repbase13 (Jurka et al. 2005) were used as the queries in BLAST searches against this local database to reveal BAC-end sequences with significant (EXPECT threshold <1.0e^{-3}) similarity matches. Characterization of tilapia sequence orthologs was semi-automated using PERL scripts. Sequences with 20% similarity or greater were assembled using GAP4 software (Staden 1.7.0; Staden et al. 2000) in 169 databases, one for each of the Repbase matching entries. The resulting 327 contigs, which had no known annotation in tilapia, were mostly annotated as follows: [repeat record number]_[frequency in the database]_[name of *D. rerio* repeat]_[matching location in this repeat]_[SRA accession number of the contig most 5′ BAC-end]. The annotated entries were constantly used to update the tilapia repeat library which was searchable using a common gateway interface web tool based on RepeatMasker (version Open 3.2.6 A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org) that we created. During contig assembly, 15 entries for repetitive sequences, which had no significant orthology to known repeats, were annotated as "unknown". To further characterize such repeats, we used MIRA software (V2.9.43; Chevreux et al. 2004) under the highly repetitive switch to assemble all BAC-ends. Output contained 27 Mb of consensus sequence in 23,722 contigs and the 528,828 filtered repeats in the file with the suffix _int_skimmarknastyrepeats_nastyseq_preassembly2.0.lst. Using Linux commands (grep, awk, sort), 332,244 unique repeats were detected. Following masking and resorting, 21,814 repeats longer than 36 bp were left. Of these, the 2,003 sequences that were longer than 200 bp were GAP4 assembled into 38 contigs that were mostly annotated as "unknown" with the indication "_MIRA" in the field used for the accession number. The rest of the repeats that were not masked by the updated library (7,452 records) were GAP4 assembled into 1,871 contigs. The 16 main contigs (repeat frequency above 60, length approximately 200 bp) were then added to the repeat library, and the rest 1,574 unique contigs with average length of 51 bp were unified using "NNNNN" spacers into one record under the entry annotation "Misc_short_repeats_

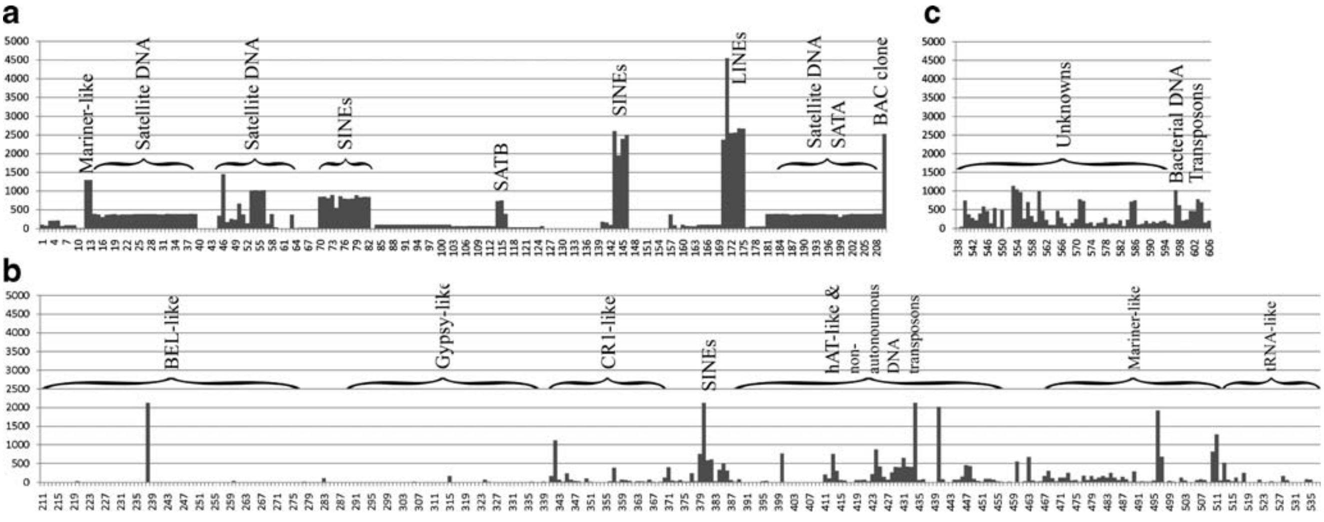

Fig. 1 Frequencies of repetitive elements in the tilapia genome. A map of the tilapia repeat library that was established from three sequence sources is shown. **a** Tilapiini repeat sequences from GenBank. **b** *Oreochromis niloticus* sequences orthologous to the repeat library of zebrafish in Repbase. **c** Novel *O. niloticus* repeats mostly detected by genome analysis using MIRA assembler. Under the *horizontal axis*, the repeat number within this library is indicated. The *vertical axis* denotes the number of significant hits obtained by BLAST search against our local database, which consist mostly of *O. niloticus* bacterial artificial chromosome-end sequences. Types of repeats that form major landmarks in this map are indicated above major frequency peaks and above regions of repeat superfamilies delineated with brackets

generated_by_MIRA". The complete repeat library is searchable and downloadable from http://cowry.agri.huji.ac.il/cgi-bin/TilapiaRM.cgi.

Results and Discussion

A total of 607 records were gathered to form the tilapia specific repeat library (Fig. 1). About a third of them (Fig. 1a) consists of entries imported directly from GenBank. To the original annotation of these records, we added a serial repeat number (RN) for the repeat in our library followed by an indication that may help to estimate the repeat frequency in the genome. This indication is the number of sequences that produced a significant alignment (EXPECT score better than 1.0e^{-3}) in a local BLAST search against the known tilapia sequences (161,071 records, approximately 120 Mbp), mostly obtained from the BAC-end sequencing project. As the RNs also reflect batch submissions and creation dates in the original databank, the distribution of peaks (Fig. 1a) is not random, and it is associated with the repeat types indicated above major landmarks on the map of the repeat library (Fig. 1). The most frequent repeat detected (Fig. 1a, RN171, frequency 4,546) was annotated as CiLINE2 repeat sequence (Oliveira et al. 1999). Indeed, L2 class of LINE-like retrotransposons from the CR1 superfamily are the most numerous repeat in Fugu (>6,500 copies; Jurka et al. 2005; Poulter et al. 1999); however, the RN171 frequency is too high to be explained by CiLINE2 alone as it was obtained from a fraction equivalent to about 11% of the genome, assuming genome size of 1,100 Mbp (Lee et al. 2005). A careful search of RN171 using Censor web tool (Kohany et al. 2006) showed that while the 5′ end contains the LINE2-like element, its 3′ end was similar to hAT-N3_FR element of the ancient and common hAT superfamily of transposons (Rubin et al. 2001). Hence, the RN171 annotation is problematic, and this chimeric element brought together the two major classes of repeats: a Class I element (retrotransposon) that moves via an RNA intermediate, and a Class II element (transposon) that migrates via a DNA intermediate. This combination was the reason for the particularly numerous BLAST hits encountered. Transposable elements nested within one another are a common situation and a known problem in repeat annotation (Kronmiller and Wise 2008). The rest of the CiLINE2 repeats (Fig. 1a, RN170–175, frequency approximately 2,500) with an estimated copy number of about 5,500 for the haploid genome of *O. niloticus* (Oliveira et al. 1999) suggest that in order to estimate the number of occurrences in the genome, the frequencies reported in this work should be at least doubled. Similar conclusion can be drawn from analysis of the frequency of 1,900 bp SATB (Fig. 1a, RN115, frequency 745), which is one of the two main satellite DNA sequences in *O. niloticus*. SATB (1,000–10,000 copies per genome) is restricted to the centromeric region of a single chromosome (Oliveira and Wright 1998). Unexpectedly, the other main satellite, the 237-bp SATA, which is distributed in the centromeric regions of all chromosomes, with tenfold higher copy number (Oliveira and Wright 1998), had fewer BLAST hits (Fig. 1a, RN206, frequency 384) than SATB.

Another notable class I repeats (Fig. 1a, RN70–82, frequency 850; RN143–146, frequency approximately 2,500) are the AFC SINEs (Terai et al. 2003), which were suggested as useful probes for the analysis of speciation of African cichlids. The most frequent group of class II repeats contained the recently identified *Misgurnus mizolepis* Tc1-like transposons (Ahn et al. 2008) of the Mariner/Tc1-superfamily (Fig. 1a, RN12–13, frequency 1,296). The identification of numerous (2,525) BLAST hits against the BAC vector prompted us to include the BAC vector sequence (RN210) at the end of the GenBank derived records (Fig. 1a, most right peak).

Most (approximately 54%) of the entries of our tilapia repeat library were derived from orthologous repeats that were present in the zebrafish repeat library in Repbase (Fig. 1b). These were used to BLAST search the BAC-end database and to form orthologous entries in our library. A few fossilized copies of the large ancient Class I repeat superfamilies with direct-orientation flanking long-terminal repeats (LTRs) of BEL (Frame et al. 2001) and of Gypsy (Britten et al. 1995) were detected (Fig. 1b, RN211–274 and RN288–337, respectively). It should be noted that the exceptional frequency of the BEL13-like element (Fig. 1b, RN238, frequency 2,132) results from chimerism of this element with a Mariner-like sequence.

The non-LTR retrotransposon of Class I repeats were more abundant (Fig. 1b, CR1-like, frequencies up to 1,129 in RN342) and have been also previously reported as widespread LINEs in teleosts (Mazzuchelli and Martins 2009). SINEs were an even more frequent non-LTR retrotransposons. Noteworthy was SINE_TE-like element (Fig. 1b, RN379–382, frequencies 588–2,131), which is a member of the V-SINE superfamily (Ogiwara et al. 2002).

A substantial portion of the high-frequency repetitive elements detected using the zebrafish Repbase library was Class II transposons (Fig. 1b). These include repeats similar to hAT and Mariner superfamilies and non-autonomous DNA transposons which rely on other active intact elements of Class II type to move them. It should be noted that in this category, chimeric repeats produced numerous BLAST hits because they contained a nested SINE within (RN434, frequency 2,125; RN496 frequency 1,929) or a combination of hAT- and Mariner-like elements (RN440, frequency 2,017). Moderate frequencies (1–174) for occurrences of repeats orthologous to tRNA pseudogenes were observed (Fig. 1b, to the right).

The third stage in assembling our tilapia repeat library was the addition of repeats that were not previously annotated or could not be detected by similarity search against GenBank or the zebrafish Repbase records (Fig. 1c). A whole genome sequence analysis to systematically detect such repeats has been recommended (Malde and Jonassen 2008). As the MIRA genome assembler (Chevreux et al. 2004) is a specialized assembler for sequencing projects with a high number of similar repeats, we took advantage of the sophisticated algorithms implemented in this assembler for disentangling repeats. These take into account the number of sequence occurrences relative to the expected coverage as well as the number of nucleotide variations within the repeated region. While assembling the available *O. niloticus* BAC-end sequences, 14% of the input sequences were annotated as repeats with an average of 3.5 repetitive sequences per read. Repetitive sequences constitute about 50% of the human genome (Tang 2007) and, consequently, its size is larger than the *O. niloticus* genome. Assuming that vertebrates have similar number of genes (Aparicio et al. 2002), it is indeed expected that the repeat content in *O. niloticus* genome would be of smaller proportion and similar to that of chicken (approximately 11% repeat content in 1,200 Mbp genome (Tang 2007)). It should be noted that although most of the genome size differences can ultimately be attributed to repeats, the precise annotation of the repeat content and the estimation of its size are complicated as there are ancient repeats that are degenerated beyond recognition.

Repeat sequences that we detected using the MIRA assembler and that were not masked by the repeat library created in the first two stages were assembled, and the consensus sequences were added to this library. A total of 59 entries that belonged to 43 groups with no significant similarity to known repeats were annotated as “unknowns”. The importance of characterizing these repeats and creating the species-specific repeat library is evident from the relative abundance of these repeats (e.g., RN554, frequency 1,044, Fig. 1c). The MIRA assembler also pointed out frequent bacterial and vector DNA contaminations that were not masked by the RepeatMasker defaults. These were added to repeat library as RN597-8 (Fig. 1c). The MIRA assembler also detected repeats that were orthologous to previously annotated transposons and escaped our analysis, as they were not present in the zebrafish Repbase library. Repbase libraries for fugu, and even invertebrates such as planaria and hydra, seem to be valuable for identifying repetitive sequences that have escaped our analysis. This work was aimed at producing a practical web tool in the form of RepeatMasker that would assist tilapia genomics. Based on *O. niloticus* repeats that our tilapia repeat masker failed to mask and that we encountered while practically using this tool, we estimate that this library represents about 80% of the repeats that would be present following similar analysis using the complete genome data.

Acknowledgements This is a contribution from the Agricultural Research Organization, Institute of Animal Science, Bet Dagan, Israel, No. 549/09. The research was supported by research grant IS-3995-07 from the USA-Israel Bi-national Agricultural Research and Development (BARD) Fund.

References

Ahn SJ, Kim MS, Jang JH, Lim SU, Lee HH (2008) MMTS, a new subfamily of Tc1-like transposons. Mol Cells 26:387–395

Andreson R, Reppo E, Kaplinski L, Remm M (2006) GENOMEMASKER package for designing unique genomic PCR primers. BMC Bioinformatics 7:172

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MDS, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJK, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297:1301–1310

Britten RJ, Mccormack TJ, Mears TL, Davidson EH (1995) Gypsy/Ty3-class retrotransposons integrated in the DNA of herring, tunicate, and echinoderms. J Mol Evol 40:13–24

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14:1147–1159

Chew JSK, Oliveira C, Wright JM, Dobson MJ (2002) Molecular and cytogenetic analysis of the telomeric (TTAGGG)(n) repetitive sequences in the Nile tilapia, *Oreochromis niloticus* (Teleostei: Cichlidae). Chromosoma 111:45–52

Ferreira IA, Martins C (2008) Physical chromosome mapping of repetitive DNA sequences in Nile tilapia *Oreochromis niloticus*: evidences for a differential distribution of repetitive elements in the sex chromosomes. Micron 39:411–418

Frame IG, Cutfield JF, Poulter RTM (2001) New BEL-like LTR-retrotransposons in *Fugu rubripes*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. Gene 263:219–230

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474

Kronmiller BA, Wise RP (2008) TEnest: automated chronological annotation and visualization of nested plant transposable elements. Plant Physiol 146:45–59

Lee BY, Lee WJ, Streelman JT, Carleton KL, Howe AE, Hulata G, Slettan A, Stern JE, Terai Y, Kocher TD (2005) A second-generation genetic linkage map of tilapia (*Oreochromis* spp.). Genetics 170:237–244

Malde K, Jonassen I (2008) Repeats and EST analysis for new organisms. BMC Genomics 9:23

Martins C, Wasko AP, Oliveira C, Porto-Foresti F, Parise-Maltempi PP, Wright JM, Foresti F (2002) Dynamics of 5S rDNA in the tilapia (*Oreochromis niloticus*) genome: repeat units, inverted sequences, pseudogenes and chromosome loci. Cytogenet Genome Res 98:78–85

Mazzuchelli J, Martins C (2009) Genomic organization of repetitive DNAs in the cichlid fish *Astronotus ocellatus*. Genetica 136:461–469

Ogiwara I, Miya M, Ohshima K, Okada N (2002) V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate Genomes and retain a strongly conserved segment within each repetitive unit. Genome Res 12:316–324

Oliveira C, Wright JM (1998) Molecular cytogenetic analysis of heterochromatin in the chromosomes of tilapia, *Oreochromis niloticus* (Teleostei: Cichlidae). Chromosome Res 6:205–211

Oliveira C, Chew JSK, Porto-Foresti F, Dobson MJ, Wright JM (1999) A LINE2 repetitive DNA sequence from the cichlid fish, *Oreochromis niloticus*: sequence analysis and chromosomal distribution. Chromosoma 108:457–468

Poulter R, Butler M, Ormandy J (1999) A LINE element from the pufferfish (fugu) *Fugu rubripes* which shows similarity to the CR1 family of non-LTR retrotransposons. Gene 227:169–179

Rubin E, Lithwick G, Levy AA (2001) Structure and evolution of the hAT transposon superfamily. Genetics 158:949–957

Smith CD, Edgar RC, Yandell MD, Smith DR, Celniker SE, Myers EW, Karpen GH (2007) Improved repeat identification and masking in Dipterans. Gene 389:1–9

Staden R, Beal KF, Bonfield JK (2000) The Staden package, 1998. Methods Mol Biol 132:115–130

Tang H (2007) Genome assembly, rearrangement, and repeats. Chem Rev 107:3391–3406

Terai Y, Takahashi K, Nishida M, Sato T, Okada N (2003) Using SINEs to probe ancient explosive speciation: “Hidden” radiation of African cichlids? Mol Biol Evol 20:924–930