Evolution of a B2 tagged sequence from a long-range repeat family in the genus Mus

Christoph Plass, Thomas Hellwig, Walther Traut, and Heinz Winking

Institut für Biologie, Medizinische Universität zu Lübeck, Ratzeburger Allee 160, D-2400 Lübeck, FRG

Received December 6, 1992; accepted January 24, 1992

Abstract. A long-range repeat family of more than 50 kb repeat size is clustered in Chromosomes (Chr) 1 of Mus musculus and M. spretus. In M. musculus this long-range repeat family shows considerable variation of copy-number frequency and contains coding regions for at least two genes. In an intron of a gene, which is part of the repeat, a B2 small interspersed repetitive element (SINE) is inserted at identical positions. The B2 element is present in all copies of the long-range repeat family; it was presumably a component of the ancestral single-copy precursor sequence that gave rise by amplification to the repeat family. Copies of the long-range repeat family vary with respect to the number of TAAA tandem repeats in the A-rich 3' end region of the B2 element. As inferred from polymerase chain reaction (PCR) data, presence and frequency of repeat number variants in the (TAAA)ₙ block are strain and species specific. The B2 element and its flanking regions were sequenced from two copies of the long-range repeat family. Sequence divergence between the two copies (only non-CG base substitutions and deletions/insertions) was determined to be 2.6%. Based on the drift rate in human Alu elements and a correction for the higher drift rates in rodents, an estimate for the divergence time of 1.7 million years was calculated. Since the long-range repeat family is present in M. musculus and M. spretus, it must have evolved by amplification before the separation of the two species about 1–4 million years ago.

Introduction

The genomes of at least two members of the genus Mus, M. musculus and M. spretus, contain a family of repeats more than 50 kb in length (long-range repeat family), which is clustered in the D band of Chr 1 (Purmann et al. 1992; Traut et al. 1992). Coding regions for two transcripts of 1.3 and 4.5 kb length are part of the repeat sequence (Eckert et al. 1991). Cloned DNA sequences from at least four different copies of the family were identified. They varied with respect to gross rearrangements, e.g., insertions and/or deletions of DNA sequences several kilobases in size, and to point mutations causing restriction site variability (Purmann et al. 1992).

The long-range repeat family is chromosomally inconspicuous in the laboratory strain C57BL, where it consists of about 50 copies per haploid genome ('HSR- genome'). In individuals from wild-mouse populations of M. musculus, variants of this cluster with higher copy numbers (for example, 800 copies) were found, which are visible as homogeneously staining regions (HSRs) in metaphase chromosomes (HSR⁺ genome, Traut et al. 1984; Purmann et al. 1992). During evolution, a progenitor sequence of the long-range repeat family was amplified to establish the low-copy repeat cluster that subsequently underwent dynamic evolution leading to high-copy-number variants (Winking et al. 1991).

Here, we investigate the sequence variability among different copies of the long-range repeat family in and around a B2 element which is a SINE (Singer 1982) inserted in an intron of the 1.3-kb RNA coding gene.

Materials and methods

Animals

The inbred M. musculus laboratory strain C57BL/6 and M. spretus, both without an HSR in Chr 1, were used as sources for HSR- DNA. The M. spretus mice were descendants from animals trapped in Porto Covo, Portugal. An HSR of a feral mouse from Mitten (Switzerland) was introduced into the C57BL/6 genome by repeated selective backcrossing. After several backcross generations the
strain was made homozygous for the HSR by inbreeding. DNA prepared from homozygous HSR carriers of this strain is termed HSR⁺ DNA.

Isolation of DNA, molecular cloning, and sequencing

Genomic, phage, and plasmid DNAs were isolated as described by Eckert and co-workers (1991). Clones Mm2, Mm8, and Mm51-Mm73 have been isolated from a genomic C57BL/6 mouse library established in λ EMBL4 (Purmann et al. 1992). Mm2K and Mm8K are EcoRI fragments of Mm2 and Mm8, respectively, subcloned in pBlueScript II SK⁺. For sequencing, the chain termination method was used (Sequencing kit, Pharmacia).

**Primers**

All mouse-specific oligonucleotide primers were synthesized with a DNA synthesizer (7500 Millipore/Milligen) in a 200-nmol scale, following the instructions given by the manufacturer. The chemicals were supplied by Millipore/Milligen. The sequences were: primer A, 21 mer: TITaCAgtctCAGcACACACTCACCTA; primer B, 17 mer: TTGTTGTAcgCAGtGAAcT; primer C, 19 mer: TCATT-GGGCAGTtCtGAG; primer D, 19 mer: GTCTGAGGTGTCTGAAGAC; primer E, 20 mer: TGCTTCAGTGGAGAGTTTGC; primer F, 23 mer: GTGCTGAGGACACGTACAGTG.

**PCR**

PCRs were set up in 100-μl assays with 5 ng of genomic DNA, 0.5 ng of cloned DNA, and 2 μl of λ phage plate lysates (about 1–5 × 10⁶ pfu/ml) as templates, 200 ng each of the appropriate primers, and 2 μl Taq polymerase (Perkin Elmer). After an initial denaturation step for 120 s at 94°C, subsequent temperature profiles were as follows: 80 s at 94°C, 120 s at 55°C, and 150 s at 72°C for 35 cycles. The last primer extension period was prolonged to 10 min.

**Calculation of sequence divergence**

All substitutions, deletions, or insertions (regardless of their length) and changes of the number of TAAA tandem repeats were counted as single events. For non-CG-related divergence, all CG-related events (NG, CN, NCG, CNG, and CGN, where N can be a substitution or a deletion/insertion) and variation in TAAA repeat numbers were omitted. Divergence was calculated by dividing the number of non-CG-related events by the total number of positions minus CG dinucleotides. For statistical evaluation of differences, the χ² test was used.

**Results**

**Mapping and sequencing of a B2 element and its flanking regions in the long-range repeat**

The two clones, Mm2K and Mm8K, represent homologous EcoRI fragments from different copies of the long-range repeat family. They were subcloned and restriction-mapped with PsI, EcoRI, HindIII, and MspI (Fig. 1). The two cloned fragments differ in an MspI restriction site present in Mm2 and missing in Mm8. Used as probes in Southern blot hybridization of genomic DNAs, they produced bands and a background smear, which indicated the presence of a highly repetitive element besides specific parts of the long-range repeat family. By its property to produce a smear, the repetitive element was mapped to the 0.3 kb PsI/MspI subfragment of Mm2K and the 0.7 kb PsI/MspI subfragment of Mm8K (Fig. 1, ‘B2’).

The 1.3 kb and 2.2 kb PsI subfragments from both clones, Mm2K and Mm8K, were subcloned, and contiguous regions of more than 600 bp (see Fig. 1, s.s.) were sequenced. The sequences are presented in Fig. 2. A stretch of 99 bp (boxed in Fig. 2) in both sequences is homologous to the published cDNA sequence of the 1.3 kb transcript (Eckert et al. 1991). It represents the last exon of the 1.3-kb RNA gene and codes for a part of the 3' untranslated region. A typical polyadenylation signal (Fig. 2, overlined) is found in this region. The two copies of the long-range repeat family represented by Mm2K and Mm8K differ from the consensus sequence of Bains and Temple-Smith (1989; Fig. 2, ‘B2’). The B2 elements in both copies of the long-range repeat family are inserted at exactly the same position in the last intron of the gene.

Compared with the consensus sequence of Bains and Temple-Smith (1989), the two B2 elements contain several substitutions, deletions, or insertions of one or two basepairs besides a prominent 10-bp deletion near the 3' end. Seven non-CG mutation hotspots had been identified in the B2 consensus sequence (Bains and Temple-Smith 1989). Our B2 sequences differ from the consensus at 3 (Mm2) and 4 (Mm8) of these positions. The internal RNA polymerase III split promoter described for B2 elements deviates from the published sequence (underlined in Fig. 2) in the first of the two promoter blocks by 4 and 3 (Mm2 and Mm8 respectively) mutations, and in the second block by 2 and 1 mutations when compared with the consensus sequence. The 3' ends of B2 elements exhibit a characteristic A-rich region, including variable numbers of TAAA simple repeats. In the two copies of the B2 element the repeat numbers are n = 4 (Mm8) and n = 7 (Mm2). The sequence differences of the two B2 elements in the long-range repeat compared with the consensus sequence are 14.7% and 15.8% (Mm8 and Mm2). Direct repeats, which normally flank B2 ele-