# Large tandem repeats in mammalian genomes *in silico* and *in situ*

Ostromyshenskii D.I., Podgornaya O.I.
I*nstitute of Cytology RAS, St. Petersburg, Russia*

## Introduction

Large tandemly repeated sequences (or satellite DNA) are necessary part of higher eukaryotes genomes and can comprise up to tens percent of the genomes. Much of TRs' functional nature in any genome remains enigmatic. TR are the most variable among different types of eukaryotic sequences up to species-specificity. The ways of TR fast evolution are not determined yet. The next generation sequencing methods and increasing number of assembled genome provide the material for the bioinformatics extracting of the nearly full set of TR in any genome. The search for the large TR lead to 62 TR's family found in mouse genome and only two of them have been known before. The aim of the current work is to compare TR sets in the genomes of closely relates species available.

## Materials and methods

Five mammalian genera was used: (1) chinese hamster *Cricetulus griseus* (Cricetudaeae, Rodentia); (2) syrian hamster *Mesocrucetus auratus*; (3) guinea pigs *Cavia* (Caviidae, Rodentia): *C. porcellus, C. apperea*; (4) bats *Myotis* (Vespertilionidae, Chiroptera): *M. brandii, M. davidii, M. lucifugus*; (5) cows *Bos* (Bovidae, Artiodactyla): *B. taurus, B. mutus, B. indicus*. Our pipeline takes into consideration the basic TR characteristic: monomer length, monomers' number in the array and the monomers' degree of diversity in the array. The methods include following steps: (1) extracting the whole TR set with TRF program ; (2) filters applied to the TR set extracted: arrays length > 3000bp, number of monomers > 4, entropy of array > 1.76; (3) nested arrays and arrays with different monomer length with similar sequences removed (4) TR set get split into families by Blast defined similarity; (5) TR families compared with Repbase to identify the known ones; (6) the resulting TR set of one species compared with the rest.

## Results

| Species | Assembly | TR % |
|---|---|---|
| *Mus musculus* | Mm_Celera | 0,122 |
| | GRCm37 | 0,026 |
| *Critulus griseus* | C_griseus_v1.0 | 0,158 |
| *Mesocricetus auratus* | MesAur1.0 | 0,1 |
| *Cavia porcellus* | Cavpor3.0 | 0,023 |
| *Cavia apperea* | CavAp1.0 | 0,013 |
| *Myotis brandtii* | ASM41265v1 | 0,084 |
| *Myotis davidii* | ASM32734v1 | 0,047 |
| *Myotis lucifugus* | Myoluc2.0 | 0,16 |
| *Bos indicus* | Bos_indicus_1.0 | 0,023 |
| *Bos mutus* | BosGru_v2.0 | 0,012 |
| *Bos taurus* | Bos_taurus_UMD_3.1.1 | 0,074 |
| | Btau_4.6.1 | 0,144 |

Table 1.The amount of large TR in mammalian genomes. Assembly indicated and large TR% in these assembly are shown. TR% counted as the ratio of all TR arrays sum to the total sequences length in current database.

| *Cavia* | *porcellus* | *apperea* | | |
|---|---|---|---|---|
| | Cpor-123 | Capp-123 | | |
| | Cpor-783 | - | | |
| | Cpor-14 | Capp-14 | | |
| | Cpor-208 | Capp-208 | | |
| | Cpor-109 | - | | |
| | - | Capp-1518 | | |
| N TR family | 26 | 10 | | |
| *Myotis* | *brandtii* | *davidii* | *lucifugus* | |
| | Mbra-258 | - | - | |
| | Mbra-17 | Mdav-20 | Mluc-381 | |
| | Mbra-80-A | Mdav-159 | - | |
| | Mbra-20 | Mdav-41 | - | |
| | Mbra-80-B | Mdav-80 | Mluc-80 | |
| | Mbra-148 | - | Mluc-154 | |
| № TR family | 133 | 105 | 26 | |
| *Bos* | *taurus* | *mutus* | *indicus* | Repbase |
| | Btau-1406 | Bmut-1402 | Bind-1406 | BTSAT4/BTAST5 |
| | Btau-1413 | - | Btau-1211 | BTSAT2/BTAST3 |
| | Btau-686 | Bmut-702 | Bind-686 | BTSAT6 |
| | Btau-48 | - | - | |
| | Btau-54 | - | - | |
| | Btau-18 | Bmut-18 | Bind-18 | |
| № TR family | 65 | 27 | 18 | |

Table 2. TR found in the assemblies indicated on table 1; in each genera the species with higher number of TR families counted as reference (1st one); top 5-6 TR are shown. TR similar in sequence (not monomer length) placed at the same line. The TR major in amount in each genome is shown in grey. Names according to Repbase for 3 known Bos TR are shown.

Genus *Cavia* (guinea pig). TR's family exist also in *C. porcellus* genome except the major TR for this species – Capp-1518. In C. porcellus genome there are two major TR – Cpor-783 is absent in the 2nd genome and Cpor-123 exists in *C. apperea* genome as the minor one.

Genus *Myotis* (bat). Only 5 TR families exist in three genome but most of TR families are species-specific. Major TR for *M. davidii* and *M. lucifugus* is common in sequence though differ in monomer length, but the same TR is minor one in *M. brandtii*. The major for *M. brandtii* is not identified in both other genomes at all.

Genus *Bos* (cow). There are three TR known for *Bos* in Repbase and all of them are found in all *Bos* assemblies. Still the major TR in all *Bos* assemblies differ: in *B. taurus* genome BTSAT4/BTSAT5 is a major TR while BTSAT6 major TR family in B. indicus genome. It is visible that most of the top TR families in genus *Bos* exist only in two genomes or even in one, i.e. is species-specific.

| Family | Genome, % | | Comments |
|---|---|---|---|
| | SRR396599_2 | SRR396609_1 | |
| 49A | 3.12 | 4.78 | MAU-BglII_M11 |
| 44A | 0.61 | 0.69 | ERV |
| 85A | 0.07 | 0.08 | L1 |
| 42A | 0.07 | 0.54 | |
| 161A | 0.07 | 0.05 | B1 |
| 62A | 0.06 | 0.05 | |
| 32A | 0.04 | 0.49 | |
| 163A | 0.03 | 0.03 | |
| 73A | 0.03 | 0.09 | |

Table3. Tandem repeats (part) in *Mesocricetus auratus* genome. Red — tandem repeats was seected for FISH mapping

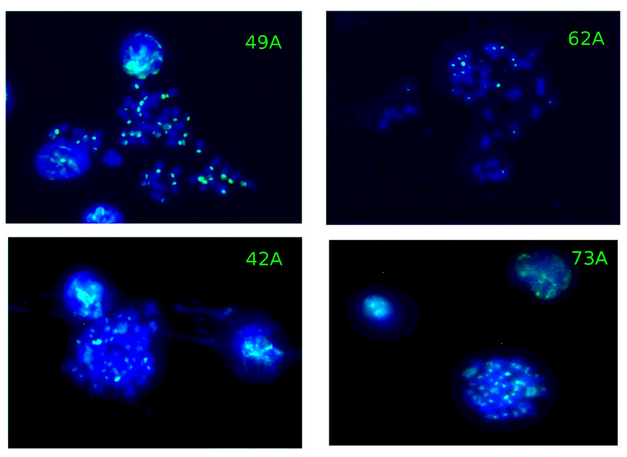


Figure 1. FISH with probe to 4 tandem repeats of *Mesocricetus auratus*

| Family | Genome content, % | | Chromosome In silico | Chromosome FISH | Comments |
|---|---|---|---|---|---|
| | SRR803182_1 | SRR803174_2 | | | |
| 49A | 0.93175 | 0.82487 | 8 | | ERV |
| 11A | 0.60389 | 0.34177 | 9-10 | | |
| 79A | 0.41075 | 0.38410 | 5,9-10,6,X | | |
| 272A | 0.34222 | 0.34247 | X | | B1 |
| 33A | 0.27568 | 0.24621 | 5 | 5,1,2 | SAU1.5 |
| 6A | 0.16142 | 0.22529 | 5,6,8-10 | | |
| 25B | 0.14970 | 0.17584 | 9-10 | | |
| 304A | 0.14572 | 0.21242 | 5 | | ERV2 |
| 77A | 0.08647 | 0.07590 | 5,2,8 | | |
| 84A | 0.07368 | 0.07757 | all | | Zn-finger |
| 17A | 0.03837 | 0.03306 | 6 | | |
| 65A | 0.03492 | 0.02827 | X | | Tc1 |
| 25A | 0.02910 | 0.02043 | 5 | 3,5,7,9-10 | |
| 27A | 0.02075 | 0.01606 | 6 | | |
| 146A | 0.01946 | 0.01103 | 5 | | |
| 62A | 0.01518 | 0.00461 | 2 | 6,2 | |

Table 4. Tandem repeats (part) in *Cricetulus griseus* genome (assembly Cgr1.0 from sorted chromosome library ). Red — tandem repeats was seected for FISH mapping.

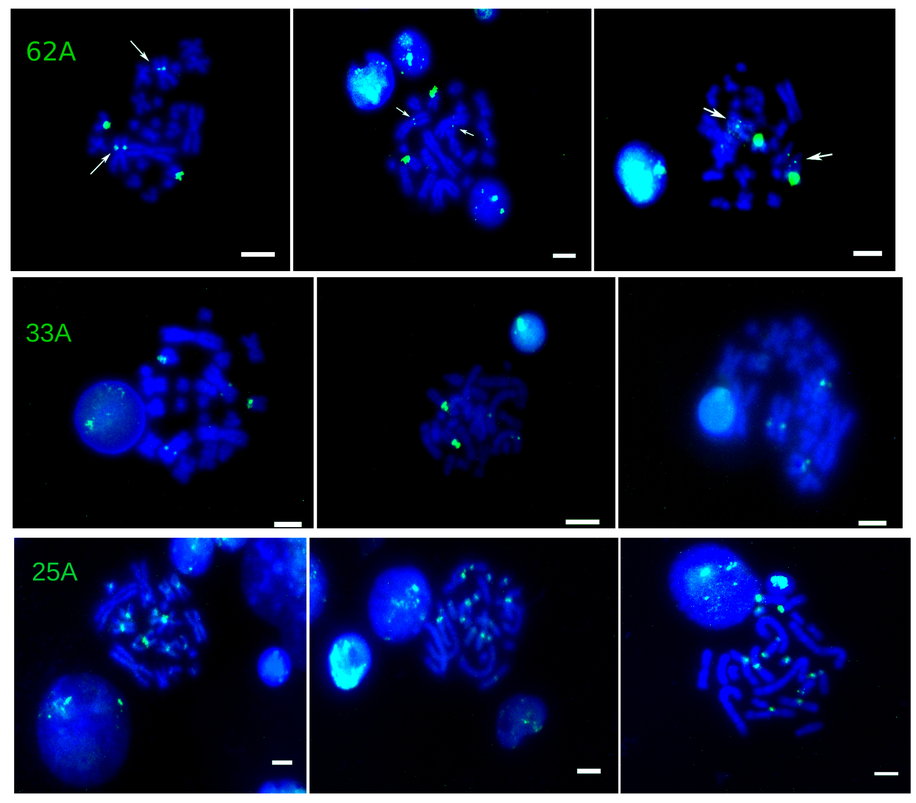


Figure 2. FISH with probe to 4 tandem repeats of *Cricetulus griseus*

## Conclusion

The absence of assembled genome of closely related species put the limitation to the bioinformatics approach. We examined all the genomes available for this aim. The most exhausting analysis of major TR (one for each species) of ~300 animals and plants display no readily apparent conserved characteristics; individual clades likely differ in terms of their tendency for closely related species to have TR that share conserved sequence characteristics (Melters et al., 2013). We compared the TR sets. Our data evidenced that there are species-specific top TR, which are absent in genome of closely related species. In all three genera examined major TRs are species-specific and hardly exist in other species of genera even as a minor ones. This finding makes the "library" hypothesis of TR evolution questionable.

Acknowledgements The work was supported by Grant 15-15-20026 Russian Science Foundation Molecular and Cellular Biology grant from the Presidium of Russian Academy of Sciences, RFBR 11-04-01700-a