The genome wide analysis of the large tandem repeats in the closely related species

Dmitrii I. Ostromyshenskii

Institute of Cytology RAS, Saint-Petersburg, Russia necroforus@gmail.com

Olga I. Podgornaya

Institute of Cytology RAS, Saint-Petersburg, Russia opodg@yahoo.com

Large tandemly repeated sequences (TR, 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 because there are only few tools available for dissecting and elucidating the TR functions. 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 wide-spread "library" hypothesis explains the occurrence of a species-specific TRs (satellite DNAs) as a result of differential amplifications within a pool of sequences shared by genomes of related species [1]. The library concept is based on comparison of TR experimentally cloned from the species of one genus, beetles *Tribolium* for example. Such an approach allows to compare a limited number of TR; the sets of the TR from the genomes was never used for comparison. *In silico* approach allows TR sets comparison. 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 [2]. The bioinformatics approach for search of only major TR in each genome have been published; unfortunately it has self-contradiction - major TR defined as centromeric, which is not true [3]. The aim of the current work is to compare TR sets in the genomes of closely relates species available.

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 [4]; (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. In the current work only 5-6 TR, top of the TR representation, is shown. TRF output analysis [2, 3] was performed with custom Python scripts.

It is known that in WGS and WGA datasets TR remain underrepresented (table 1, [2]). It is visible that TR amount vary ~0.1-0.01%, which is far less than the experimentally determined amount of the mouse Major Satellite (MaSat) alone (~8%) [2]. TR amount of two *M.musculus* genomes assemblies, already checked, are in this interval (table 1), still the tiny TR representation in dataset reflect the TR families representation in the genome [5].

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.

Genus *Mus* was the 1st to compare due to the *M.musculus* genome mostly well investigated in the sense of TR content. We tried to find all the 62 *M.musculus* TR families [2] in raw reads of *M. caroli* genome (Caroli Genome Project, PRJEB2188). There are only few TR of *M. musculus* in *M. caroli* genome. *M. musculus* major satellite (MaSat or GSAT-MM) occupied nearly 0,7% of *M. caroli* genome, while in *M. musculus* genome - ~ 11 %. In *M. caroli* genome we found 5 other *M. musculus*'s TR's families (table 2). The amount of 4 of them are of the same order with the exception of TR-1590-A-MM, which belongs to the class of transposable element (TE) related TR [2]. Two known as previously cloned *M. caroli* TR family [6] are not found in *M. musculus* genome [5].

Family	*Mus musculus*	*Mus caroli*
MaSat	11.3	0.66
TR-1590A-MM	2.24	0.04
TR-6A-MM	0.25	0.15
TR-31B-MM	0.22	0.19
TR-107A-MM	0.04	0.04
TR-57A-MM	0.004	0.044

Table 2. M.musculus TRs (family) representation in both genome in %. Calculation made with alignment of raw reads to TR arrays. Methods details in [5].

The TR families found in the next three genera mostly absent in Repbase. So the TR nomenclature consists of short species name and monomer length in bp (table 3).

Cavia	*porcellus*	*apperea*		
	Cpor-123	Capp-123		
	Cpor-783	-		
	Cpor-14	Capp-14		
	Cpor-208	Capp-208		
	Cpor-109	-		
	-	Capp-1518		
Numb. 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	
Numb. 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	
Numb. TR family	65	27	18	

Table 3. 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). *C. porcellus* genome possesses 25 TR and *C. apperea* – only 10 TR. 9 out of 10 *C. apperea* 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). There is no any TR of *Myotis* in Repbase, but 133 TR's families are found in *M. brandtii* genome, 105 - in *M. davidii* genome and 26 - in *M. lucifugus* genome. 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.

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 (table 1-3). 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 [3]. 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.

Acknowledgments. This work was supported by the Russian Foundation for Basic Research (13-04-01739-a) and grant from presidium RAS (MCB).

1. Plohl, M., Meštrović, N., Mravinac, B. (2014). *Chromosoma*, 123(4), 313-325.
2. Komissarov, A. S. et al (2011). *BMC genomics*, 12(1), 531.
3. Melters, D. P. et al. (2013). *Genome Biol*, 14(1), R10.
4. Benson G. (1999) *Nucleic Acids Research* Vol. 27, No. 2, pp. 573-580.
5. Ostromyshenskii, D. I. et al (2015). *Tsitologiya*, 57(2), 102-110.
6. Kipling D et al (1995) *Mol Cell Biol* 15:4009–4020