doi 10.18699/bgrs2024-1.1-25

Investigation of large tandem repeats in the genome of Grass frog (*Rana temporaria*)

Popova M.[1, 2]*, Komissarov A.[3], Podgornaya O.[2], Travina A.[2]

[1] *Center for Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Moscow, Russia*

[2] *Institute of Cytology, RAS, St. Petersburg, Russia*

[3] *Applied Genomics Laboratory, SCAMT Institute, ITMO University, St. Petersburg, Russia*

* *marinaalexpopova@yandex.ru*

Key words: tandem repeats; *Rana temporaria*; repetitive sequences

Motivation and Aim: Non-coding repetitive DNA sequences, particularly transposable elements (TEs) and tandem repeats (TRs), form significant parts of eukaryotic genomes. TRs are crucial for chromosomal functions such as kinetochore formation and genome organization. Despite their importance, TRs remain understudied, especially in non-mammalian species. Amphibians, with their large and variable genome sizes, serve as excellent models for studying TRs. The grass frog, *Rana temporaria*, occurs in Europe, including Russia, and is an ideal subject due to its sequenced genome and potential for experimental validation of computational predictions. This study aims to identify and map TRs in the *Rana temporaria* genome using bioinformatics and experimental approaches, including *in situ* hybridization, to enhance our understanding of the molecular mechanisms governing development and evolution in amphibians and contribute to the broader field of genomic architecture in vertebrates.

Methods and Algorithms: Publicly available Illumina reads from the *Rana temporaria* low-coverage genome sequence project (PRJNA294436) and genome assembly (aRanTem1.1 version from NCBI, GCF_905171775.1) were utilized. K-mers (k=23) were extracted from the raw reads after removing Illumina adapters, sorted by frequency, and assembled into circular paths using a de Bruijn graph approach with the extracTR tool. The most frequent k-mers in each path were identified and used to create labeled oligonucleotide probes for fluorescence *in situ* hybridization (FISH). To confirm the extracTR results, TAREAN was also employed on a one-million-read subset. Tandem Repeat Finder (TRF) was used to identify TRs in the genome assembly, focusing on those over 10,000 bp and excluding those within protein-coding genes. Monomers of identified TRs were converted into dimer form for further analysis. A blastn search was used to cluster repeats into families, and their similarity to transposable elements was assessed using the DFAM database. Results were visualized with Python's plotly library.

Results: Using extracTR, we identified six abundant TRs from reads, comparing these with TAREAN results. Two k-mer families detected by extracTR were not identified by TAREAN, and the reads per million estimate for one specific family (Rtem494A, also known as s1a) was lower in extracTR than in TAREAN, likely due to TAREAN's input data limitations. Six probes targeting the most abundant TRs in raw reads were used for *in situ* hybridization on chromosomes, with probes derived from the most conserved monomer sequences, except for TR Rtem494a, for which the additional PCR-amplified probe also was used due to its longer monomers. The *R. temporaria* karyotype, with 26 chromosomes, showed intense hybridization signals mainly at the centromeres of

large and some small chromosomes. Additional signals were noted on chromosome arms, particularly with Rtem47a. Rtem219a probes, the latter correlating with the non-transcribed spacer of 5S rDNA and avoiding the nucleolus organizer region on chromosome 10, displays intense signals in the centromeric regions of 5 big and 2 small chromosomes. Signal from centromere of chromosome 7 is probably to coincide to the only 5S rDNA loci, found a in the early studies [1].

Using the aRanTem1.1 reference genome, 768 tandem repeat arrays exceeding 10,000 bp were identified, grouped into 76 families or classified as singletons, with many located on unplaced or unlocalized scaffolds. The identified TR families varied in monomer lengths (4–1930 bp) and GC content (27.6–65.95 %). Four out of the six most abundant TRs in raw reads were also prominent in the genome assembly. Dfam database searches revealed no significant similarity between TR family consensus sequences and transposable elements, though some TRs showed likeness to tRNA and snRNA genes. Chromosome specificity varied among TR families, with several located on multiple or unplaced scaffolds. Despite the low representation of TRs in the reference genome (0.32 % for TRs over 10 kb), their distribution along chromosome arms and enrichment in pericentromeric and subtelomeric regions were mapped, revealing some discrepancies between *in situ* and *in silico* data, highlighting challenges in repetitive DNA assembly.

Two noteworthy findings were: (1) TRs of the Rtem219A family consist of 5S rRNA sequences and a non-transcribed spacer, and (2) probe Rtem47A, derived from raw reads, displayed strong FISH signals but was found in only a few copies across three specific TR arrays in the assembly, suggesting high repetitiveness across the genome. We hypothesize that probe Rtem47A represents a dispersed repeat containing an inner tandem repeat.

Conclusion: Repetitive DNA sequences, including TRs, are challenging to assemble and annotate due to their complex structures and prevalence in centromeric, pericentromeric, and subtelomeric regions, which are typically underrepresented in genome assemblies. All TRs found with TRF constitute about 10.23 % of the *R. temporaria* genome, with many arrays located on contigs with unresolved gaps. Assembly-free methods like TAREAN detect abundant TRs but struggle with low-copy and short dispersed TRs. Using a more extensive dataset with the extracTR tool allowed us to identify TRs that TAREAN missed. FISH successfully detected all six targeted TRs on *R. temporaria* chromosomes, validating the effectiveness of these bioinformatics approaches.

The Rtem219a-probe hybridization in the centromeric regions of several chromosomes indicates the evolutionary dynamism of rDNA, with potential spread across the genome creating new loci, non-functional pseudogenes, and possibly satellite DNA (TRs), but further detailed studies are required to clarify the nature and functionality of these 5S rRNA gene variants and pseudogenes. Such TRs, arising from 5S rDNA, have already been described for some species of plants, fish and amphibians. This underscores the complexity and evolutionary significance of rDNA organization in the genome.

Funding: The study is supported by a grant from the Russian Science Foundation (No. 24-24-00480).

References

1. Vitelli L., Batistoni R., Andronico F., Nardi I., Barsacchi-Pilone G. Chromosomal localization of 18S+28S and 5S Ribosomal RNA genes in evolutionarily diverse anuran amphibians. *Chromosoma*. 1982;84(4):475-491. doi 10.1007/BF00292849