Tandem repeat motif characterization for precision medicine: A brief overview of conventional methods and massive parallel sequencing approaches

Ingrid Lojova[1,2,3], Jaroslav Budis[2,3,4,5], Tomas Szemes[2,3,4], Monika Buchalova[2,6], Jan Radvanszky[1,2,3]

[1]Institute of Clinical and Translational Research, Biomedical Research Center of the Slovak Academy of Sciences, Bratislava, Slovakia
[2]Comenius University Science Park, Bratislava, Slovakia
[3]Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia
[4]Geneton Ltd., Bratislava, Slovakia
[5]Slovak Center of Scientific and Technical Information, Bratislava, Slovakia
[6]Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia

Repetitive motifs, also known as tandem repeats, are stretches of DNA that are repeated multiple times in a row within a genome. These motifs have important functional and clinical implications in human genomes, including their involvement in genetic disorders such as Huntington's disease, Fragile X syndrome or myotonic dystrophies. Reliable and accurate characterization of these motifs, especially on genome scales, has remained a challenge due to their structural complexity and variability. In this review, we explore the possibilities of various molecular methods for the reliable characterization of clinically-relevant repetitive motifs, including molecular biology techniques, high-throughput sequencing technologies and bioinformatics tools.
Keywords: tandem repeats, conventional TR genotyping, massively parallel sequencing

Charakterizácia tandemových repetitívnych motívov v precíznej medicíne: Stručný prehľad konvenčných metód a masívne paralelného sekvenovania
Repetitívne motívy, známe aj ako tandemové opakovania, sú úseky DNA, ktoré sa v genóme opakujú viackrát za sebou. Tieto motívy majú v ľudských genómoch dôležitý funkčný a klinický význam vrátane ich účasti na genetických poruchách, ako je napríklad Huntingtonova choroba, syndróm fragilného X alebo myotonické dystrofie. Spoľahlivá a presná charakterizácia týchto motívov, najmä na úrovni genómu, predstavuje vzhľadom na ich štrukturálnu zložitosť a variabilitu veľkú výzvu. V tomto prehľade sumarizujeme možnosti rôznych molekulárnych metód na spoľahlivú charakterizáciu klinicky relevantných repetitívnych motívov vrátane klasických metód molekulárnej biológie, vysokoparalelných sekvenčných technológií a bioinformatických nástrojov.
Kľúčové slová: tandemové opakovania, konvenčná TR genotypizácia, masívne paralelné sekvenovanie

Newslab, 2023; roč. 14 (1): 40 – 44

Introduction

Tandem repeats (TRs) are stretches of DNA sequences, which contain repeated units several times. The sequence composition and length of the repeating units may be various in different TRs, and they can be found in various locations in the human genome, including in intergenic regions, introns, and exons of genes. Typically the number of repetitions is highly variable between individuals. Since they represent extremely variable genomic regions, TRs have many important biological roles in several physiological and pathophysiological processes. Their proper characterization may, therefore, explain observed phenotypes of traits in biomedical context, and thus markedly improve informativeness of clinical genomics assays. They also became utilized in several non-biomedical applications, mainly in population genetics, forensic as well as genealogical applications[(1)].

It is generally known and well described, that the number of repeat units in certain TRs causes a pathogenic effect when it exceeds certain thresholds. Several studies have localized more than 50 TRs on the genome, where such pathogenic expansions cause severe neurological, neurodegenerative and neuromuscular disorders, conventionally known as repeat expansion disorders (REDs)[(2)]. These are, in general, monogenic diseases, although their inheritance is not always fully mendelistic and anticipation is commonly described in such families (e.g. Huntington´s disease, Fragile X syndrome or Myotonic dystrophies). In addition, reports started to

appear which describe the potential role of TRs in the determination of complex phenotypes[3], and very recently, the first genome-wide association studies of TRs were published, for example, for Parkinson's disease[4]. This clearly outlines the application potential of TR characterisation not only in the field of diagnostics and complex disease biology, but also in the possible preventive care against several common complex diseases, through potentially more effective identification of at-risk individuals.

Detailed and accurate characterization of TR motifs, in terms of complexity of either structures or biological effects, is an absolutely crucial aspect in assessing their clinical relevance, creating a major challenge for methods aimed at their genotyping. The main challenges of accurate TR characterization include the following factors: 1) The resistance of long TR motifs to amplification by PCR; 2) The complexity of some TR regions, in which interruptions of otherwise pure motifs may occur, or interruptions may be lost, or the clinically relevant motifs may be part of more complex repeat structures; 3) The stutter effect, caused by typical errors of DNA polymerases in TR regions, which makes it difficult to identify the correct fragment in a set of multiple non-specific fragments; and 4) The errors made by classical mapping tools used in sequencing analyses (designed to align sequencing reads containing single nucleotide variants or small insertion-deletions), which are not adapted to the high variability of TR sequences. Therefore, the success of TR analysis highly depends on several factors, including the length and complexity of the repeat sequence and the availability of specific laboratory technologies. Here we provide a basic overview of the most commonly used methods for this purpose. For this review, we decided to split them into two major categories, i.e. those offering more-or-less detailed characterisation of individual loci, with limited possibilities of multiplexing, and those based on sequencing, especially massively parallel sequencing (MPS), offering genome-wide high-throughput characterisation of TR motifs.

Conventional low-throughput TR genotyping methods

For decades, conventional molecular-biology techniques have been used to identify tandem repeat variations, which may be particularly helpful in more effective prediction and/or diagnosis of many human diseases, especially those having monogenic phenotype determination. From these, PCR modifications, and in certain cases also Southern blotting, are still the most commonly used methods, which are considered gold standard in DNA based laboratory diagnostics of TR associated diseases.

Southern blotting

The principle of the method is based on cleavage of DNA fragments by restriction enzymes, which are subsequently separated by gel electrophoresis. The next step is to transfer the DNA fragments to a carrier membrane (usually nylon or nitrocellulose) followed by detection of the target DNA fragment by hybridisation with a labeled probe specific for the TR region. The number of repeats can be determined based on the size of the fragment that hybridizes to the probe. Since it does not require PCR amplification, Southern blot allows to determine the size of long, pathogenically expanded TR loci. On the other hand, it is not suitable for the differentiation of small alleles, especially those in the healthy and premutational range[5]. The limitations of this method lie mainly in its time-consuming and labor-intensive nature, technical difficulty, and it also requires large amounts of intact high-molecular-weight genomic DNA.

PCR-based methods

The most commonly used technology for amplification and characterisation of TR loci is polymerase chain reaction (PCR), followed by different options of evaluation of amplified fragments. Historically, several technologies have been used, such as evaluation based on restriction fragment length polymorphism (RFLP)[6], on denaturing high-performance liquid chromatography (dHPLC)[7], or even based on the denaturation characteristics and melting temperatures of the amplified fragments using high-resolution melting analysis (HRM)[8]. Historically, fragment size analysis by agarose or polyacrylamide gels (PAGE) are well known approaches, however, the most commonly used are automated capillary electrophoresis based separation techniques using genetic analyzers. Agarose electrophoresis allows efficient separation of PCR products only for alleles with a sufficiently large difference or in the case of large alleles. PAGE has a much higher resolution and can separate even fragments that differ by only a few base pairs. Capillary electrophoresis represents the most efficient platform to analyze length profiles of fluorescently labeled PCR products. The advantage of the methodology is that it is both affordable and not specifically labor-intensive, can be automated, and also multiplexed for certain extents[9,10]. Conventional PCR is mainly used to detect alleles belonging to normal ranges or premutation lengths, up to around 300 - 600 nucleotides. For longer alleles, i.e. larger repeats, this method is inaccurate and fails. On the other hand, long-range PCR protocols have been described too. Long-range PCR is used to characterize TRs that exceed the detection capability of conventional PCR. This methodology uses high-fidelity polymerase and optimized reaction conditions that include extended annealing time and elevated annealing temperature. Such altered reaction conditions allow an extension of the effective range of allele detection, but still only up to a certain size. For visualization of larger alleles, hybridization-based detection may be required too[11]. Another specific modification of conventional PCR is the so-called small-pool PCR that was described and used to resolve discrete bands in samples giving smeared or diffused bands by Southern blotting. This method is based on partial elimination of preferential amplification and amplification of too many fragments of different length. This is achieved by diluting and separating the template molecules into different individual amplifications. It also allows to increase the range of detected alleles, however, again only up to certain lengths limited by the PCR capabilities themselves[12]. Beyond the well known length limitations, conventional PCR based methods are limited also by the possibilities of getting false results due to: 1) „*stutter effect*" caused by DNA polymerase errors; 2) migration changes due to sequence interruptions[13]; or even by 3) allelic dropout due to sequence changes under the primer binding sites.

A specific and highly effective method, called repeat-primed PCR (RP PCR), was developed to overcome the main limitation of conventional PCR (and also of the majority of its modifications which use primer pairs flanking the region of interest), i.e. the inability to amplify through large and complex motifs. It uses two (flanking primer and repeat-specific primer) or three primers (flanking primer, repeat-specific primer and tail primer) per reaction. Flanking primer is fluorescently labeled and it is localized upstream or downstream to the repetitive region. The second primer (repeat-specific) is designed directly in the region of the repetitive motif, generating multiple products. The largest of these products reflect the size of the entire allele, however, with certain limitations. The reactions may be made more specific if a third (tail) primer is used that has no complementarity to the human genome. However, this primer is present in limited amounts in the reaction, resulting in a third primer taking over after a few PCR cycles. This third primer targets the PCR products that were produced by the previous two primers and amplifies them all, generally in a way, in which the repeat specific primer is modified by a 5´ tail that creates a template for the third primer. The evaluation is also combined with automated capillary electrophoresis. RP PCR is able to point out single locus expansions[14,15], however, multiplex versions were also described[9,16]. The disadvantage is that it is unable to determine the specific length of the expanded allele and, for certain complex loci, it was determined to be not 100 % sensitive[17,18]. In addition to expansion detection, RP-PCR is able to identify the presence of sequence interruptions by other motifs within the base repetition. Sequence interruptions, on the other hand, may also disrupt RP-PCR signals to such an extent that they can result in false negative conclusions[13]. Similarly to conventional PCR, false negative results can also occur in a form of allelic dropout, due to sequence variants in the primer binding site. For the mentioned limitations, bidirectional RP-PCR, performed from both ends of the repeat motif, are generally offered to increase sensitivity[9,13,17].

DNA sequencing methods

Following the introduction and spreading of commercial automatized genetic analyzers, direct DNA sequencing allowing the characterization of the primary structure of DNA, became established as a standard method for laboratory diagnostics and remained in this position for decades. However, in the field of TRs characterisation it had no special role, until second generation sequencing platforms began to appear. This, however, required also specific bioinformatic tools to be developed, which are dedicated to TRs genotyping from millions of short sequencing reads.

Sanger sequencing

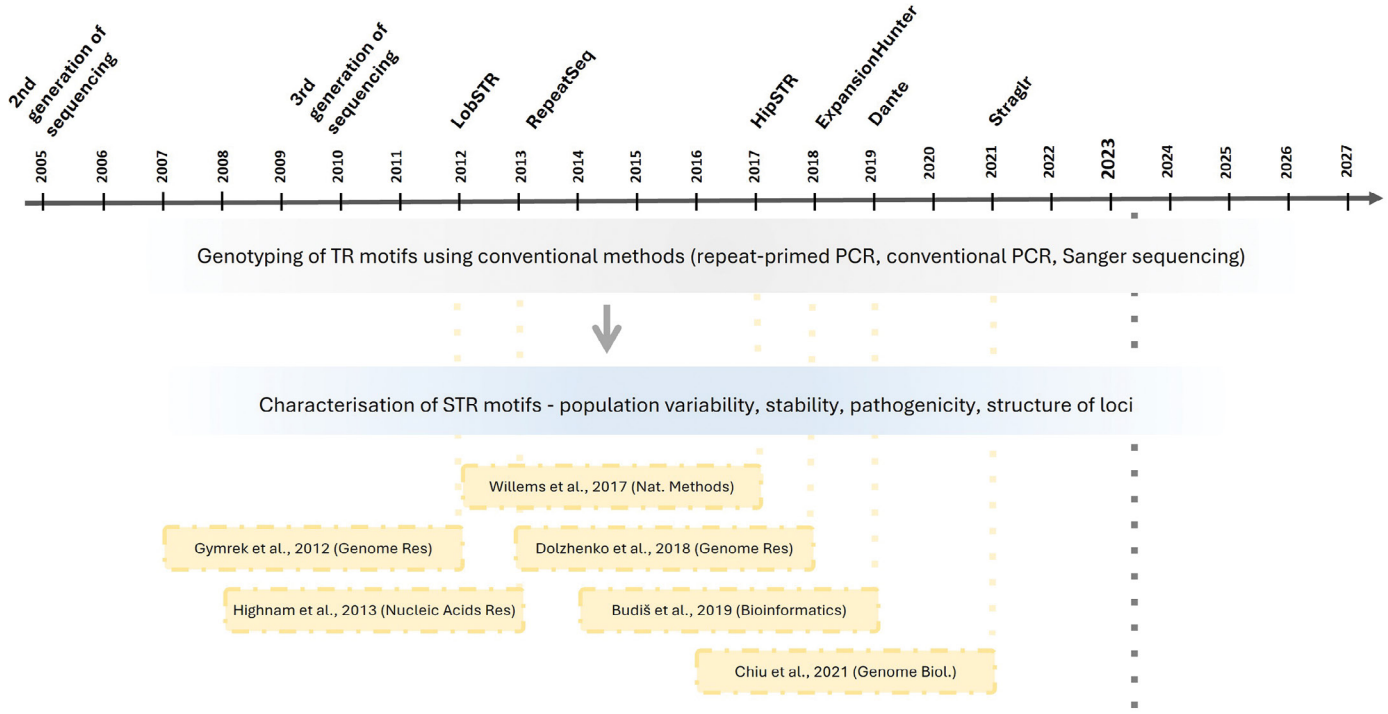
Sanger sequencing represents the most widespread first generation sequencing method and was considered as a gold standard among molecular diagnostic techniques allowing screening and genotyping of several types of sequence variants. For genotyping of TR loci, however, it was not routinely used because of several reasons. The disadvantage of the methodology lies in the detection limit (up to approximately 1000 nucleotides). Another problem lies in the analysis of alleles with a heterozygous constitution, in which based on the shifted signal, a larger allele cannot be determined[19]. On the other hand, it can be used, for example, during standardization of other molecular methods, such as conventional and repeat-primed PCR for the fragment sizing. In addition, it was also described to be useful in more detailed characterisation of repeat structures, including sequence interruptions[20,21].

Massively parallel sequencing

Within MPS applications, we distinguish at least two generations of sequencing methods including those having typically relatively short sequencing reads and requiring clonal amplification before sequencing (second generation) and those allowing longer reads and allowing single molecule sequencing, i.e. which does not require clonal amplification (third generation). From the technical point of view, when considering for example the target genomic region, sequencing can be aimed at the whole genome (WGS), but also can be narrowed to its certain parts, such as to smaller gene panels, larger panels, or even whole exomes (WES). Another relevant technical feature is, whether the preparation of the sequencing library is associated with PCR amplifications (may introduce errors) or are PCR-free (allows reduced error rate and is typically available for WGS). In addition to laboratory and technical aspects, bioinformatic processing of sequencing data may also be challenging and highly variable, while still under intensive development and diversification. In general, actually the most widespread, especially in a routine clinical setting, are platforms belonging to the second generation of sequencing supporting short-read sequencing (especially platforms of the Illumina company). Although still mainly in research settings, some third generation platforms (such as the PacBio or nanopore based sequencing) are appearing as potential tools for more effective TRs characterisation, especially when considering the possibilities of detection of large expansions, or the characterisation of highly repetitive regions. Possibilities and limitations of TR detection are largely connected to the chosen sequencing generation, platform, technical details of library preparation as well as on the applied bioinformatic pipelines.

Second generation platforms generally rely on sequencing following fragmentation of the DNA into small pieces (alternatively on amplicon sequencing), attaching adapters to the ends of the fragments, and then amplifying and sequencing the fragments using a next-generation sequencing instrument. These platforms are generally highly accurate and allow sequencing of short reads of DNA (up to 2 x 300 base pairs, yet). These approaches allow high-throughput characterisation of TRs throughout the genome, although with this regard the main limitations are clear if WES or smaller panels are used, i.e. the TR loci of interest need to be covered by reads. Another limitation is represented by the necessity of PCR amplification, each amplification step introduces challenges and possible errors to TR genotyping. Another limitation lies in the length of the TR regions, since to detailed characterisation they need to be covered by reads in their total length with certain unique flanking regions on their both ends[22,23]. The presence of repeat expansions, even the largest ones, may be indirectly identified from different read characteristics even if they exceed the sizing limit of the re-

Figure1. *Timeline of the evolution of TR locus characterization. Initially, it was possible to genotype TR motifs mainly by methods that represent the gold standard in molecular diagnostics (classical PCR, repeat-primed PCR, Sanger sequencing, Southern blotting,...), but they are still used as validation methods. After the spread of second-generation sequencing platforms into mainstream laboratories, several sophisticated tools have been developed over the last decade specifically for the specific nature of TR motifs (LobSTR, RepeatSeq, HipSTR) but also for the detection of expansions (Expansion Hunter, Dante, Straglr) and the number of these tools is still growing.*



ads (i.e. even if they have no reads sequencing through the entire motif), such as the presence of partial reads (reads containing one of the flanking regions and part of the repeat region) or the position of read-pairs, if paired and sequencing was used[24,25].

On the other hand, larger read lengths and overcoming amplification to a larger extent allows the use of third generation sequencing platforms for more effective TR characterisation, especially by allowing larger and more complex repeats to be directly sequenced by individual spanning reads, even of those belonging to larger expansions. This method makes it possible therefore to identify larger TRs that may be missed by shorter read lengths. These sequencing platforms have, however, some limitations too, including a relatively high error rate and a relatively high cost per base compared to other sequencing technologies. They are therefore most often used in combination with other sequencing technologies, such as those offered by second generation platforms, to generate high-quality, high-resolution genomic data.

Bioinformatic tools

TR typing at the genomic level has become possible with some delay and took longer to catch on, mainly due to certain technical limitations of the conventionally used bioinformatics tools processing MPS data, and high variability in individual TR loci. The majority of genotyping tools are not adapted to TR sequence variability, and they tend to „lose" some reads in certain specific cases[26]. In such genotypes, mapping often fails, which can lead to the loss of reads and incorrect clinical interpretation of predicted genotypes. Several more sophisticated tools have been lately designed specifically for the specific nature of the TR motifs, or even for expansion detection. These, such as STRetch, ExpansionHunter, Dante[24,27,28], use complex statistical modeling to estimate the underlying genotypes of TR motifs influenced by the stutter effect (**Figure 1**).

Conclusion

Tandem repeat genotyping is a challenging task due to several inherent characteristics of these DNA sequences. These include the repetitive nature itself, creating challenging sequences for amplification, different and highly variable lengths of repeating motifs, the possible presence or loss of sequence interruptions, but also the complexity of motif structures. Currently, TR motifs are most commonly characterized, in routine practice, by conventional, low-throughput methods. Within these, conventional PCR and repeat-primed PCR are likely the most commonly used. Other PCR modifications, such as long-range PCR or small-pool PCR, or even Southern blotting, are still used, however, they have some specificities and limitations. On the other hand, progreding genomic technologies, such as MPS applications, started to revolutionize genomics and disease diagnostics. The main focus in current development is, therefore, on the possibilities of characterization of TR loci by MPS applications and on improving the accuracy and reliability of TRs genotyping by more effective integration of molecular biology and bioinformatics techniques.

Acknowledgment
This publication was created with the support of the following resources: Operational Program Integrated Infrastructure for the projects: Center for Biomedical Research – BIOMEDIRES – II. Phase, ITMS: 313011W428; PanClinCov, ITMS: 313011ATL7, co-financed by the European Regional Development Fund; Slovak Research and Development Agency under the project ID APVV-18-0319; and Scientific Grant Agency under the project ID VEGA_2/0146/23.

Poďakovanie
Táto publikácia vznikla s podporou týchto zdrojov: Operačný program Integrovaná infraštruktúra pre projekty: Centrum pre biomedicínsky výskum - BIOMEDIRES – Fáza II., kód ITMS: 313011W428; PanClinCov, kód ITMS: 313011ATL7, spolufinancované zo zdrojov Európskeho fondu regionálneho rozvoja; Agentúra na podporu výskumu a vývoja SR v rámci projektu ID APVV-18-0319 a Vedecká grantová agentúra v rámci projektu ID VEGA_2/0146/23.

REFERENCES

1. Balzano E, Pelliccia F, Giunta S. Genome (in) stability at tandem repeats. Seminars in Cell & Developmental Biology 2021; Vol. 113: 97-112.
2. Malik I, Kelley CP, Wang ET, et al. Molecular mechanisms underlying nucleotide repeat expansion disorders. Nature Reviews Molecular Cell Biology 2021; 22(9): 589-607.
3. Gymrek M, Goren A. Missing heritability may be hiding in repeats. Science 2021; 373(6562): 1440-1441.
4. Bustos BI, Billingsley K, Blauwendraat C, et al. Genome-wide contribution of common short-tandem repeats to Parkinson's disease genetic risk. Brain 2023; 146(1): 65-74.
5. Buxton J, Shelbourne P, Davies J, et al. Detection of an unstable fragment of DNA specific to individuals with myo-tonic dystrophy. Nature 1992; 355: 547-548.
6. Takahashi S, Miyamoto A, Oki J, et al. CTG trinucleotide repeat length and clinical expression in a family with myotonic dystrophy. Brain and Development 1996; 18(2): 127-130.
7. Oefner PJ. et Underhill PA. DNA mutation detection using denaturing high-performance liquid chromatography (DHPLC). Current protocols in human genetics 1998; 19(1): 7-10.
8. Nicklas JA, Noreault-Conti T, Buel E. Development of a fast, simple profiling method for sample screening using high resolution melting (HRM) of STRs. Journal of forensic sciences 2012; 57(2): 478-488.
9. Radvansky J, Ficek A, Kadasi L. Upgrading molecular diagnostics of myotonic dystrophies: multiplexing for simultaneous characterization of the DMPK and ZNF9 repeat motifs. Molecular and Cellular Probes 2011; 25(4): 182-185.
10. Bauer PO, Kotliarova SE, Matoska V, et al. Fluorescent multiplex PCR: fast method for autosomal dominant spinocerebellar ataxias screening. Russian Journal of Genetics 2005; 41: 675-682.
11. Cheng S, Barcelo JM, Korneluk RG. Characterization of large CTG repeat expansions in myotonic dystrophy alleles using PCR. Hum Mutat 1996; 7: 304-310.
12. Wong LJ, Ashizawa T, Monckton DG, et al. Somatic heterogeneity of the CTG repeat in myotonic dystrophy is age and size dependent. Am J Hum Genet 1995; 56: 114-122.
13. Radvansky J, Ficek A, Minarik G, et al. Effect of unexpected sequence interruptions to conventional PCR and repeat primed PCR in myotonic dystrophy type 1 testing. Diagnostic Molecular Pathology 2011; 20(1): 48-51.
14. Warner JP, Barron LH, Goudie D, et al. A general method for the detection of large CAG repeat expansions by fluorescent PCR. J Med Genet 1996; 33: 1022-1026.
15. Falk M, Vojtiskova M, Lukas Z, et al. Simple procedure for automatic detection of unstable alleles in the myotonic dystrophy and Huntington's disease loci. Genet Test 2006; 10: 85-97.
16. Lian M, Limwongse C, Yoon CS, et al. Single-tube screen for rapid detection of repeat expansions in seven common spinocerebellar ataxias. Clinical Chemistry 2022; 68(6): 794-802.
17. Radvansky J, Ficek A, Kadasi L. Repeat-primed polymerase chain reaction in myotonic dystrophy type 2 testing. Genetic Testing and Molecular Biomarkers 2011; 15(3): 133-136.
18. Catalli C, Morgante A, Iraci R, et al. Validation of sensitivity and specificity of tetraplet-primed PCR (TP-PCR) in the molecular diagnosis of myotonic dystrophy type 2 (DM2). J Mol Diagn 2010; 12: 601-606.
19. Gettings KB, Kiesler KM, Faith SA, et al. Sequence variation of 22 autosomal STR loci detected by next generation sequencing. Forensic Science International: Genetics 2016; 21: 15-21.
20. Musova Z, Mazanec R, Krepelova, A, et al. Highly unstable sequence interruptions of the CTG repeat in the myotonic dystrophy gene. American journal of medical genetics Part A 2009; 149(7): 1365-1374.
21. Braida C, Stefanatos RK, Adam B, et al. Variant CCG and GGC repeats within the CTG expansion dramatically modify mutational dynamics and likely contribute toward unusual symptoms in some myotonic dystrophy type 1 patients. Human molecular genetics 2010; 19(8): 1399-1412.
22. Gymrek M, Golan D, Rosset S, et al. lobSTR: a short tandem repeat profiler for personal genomes. Genome research 2012; 22(6): 1154-1162.
23. Highnam G, Franck C, Martin A, et al. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic acids research 2013; 41(1): e32-e32.
24. Budiš J, Kucharík M, Ďuriš F, et al. Dante: genotyping of known complex and expanded short tandem repeats. Bioinformatics 2019; 35(8): 1310-1317.
25. Dolzhenko E, Van Vugt JJ, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome research 2017; 27(11): 1895-1903.
26. Willems T, Zielinski D, Yuan J, et al. Genome-wide profiling of heritable and de novo STR variations. Nature methods 2017; 14(6): 590-592.
27. Dashnow H, Lek M, Phipson B, et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome biology 2018; 19(1): 1-13.
28. Tankard RM, Bennett MF, Degorski P, et al. Detecting expansions of tandem repeats in cohorts sequenced with short-read sequencing data. The American Journal of Human Genetics 2018; 103(6): 858-873.

Mgr. Ingrid Lojová
Institute of Clinical and Translational Research,
Biomedical Research Center of the Slovak Academy of Sciences
Dúbravská cesta 9, 845 05 Bratislava, Slovakia
e-mail:ingrid.lojova@gmail.com