UDC 575.
https://doi.org/10.2298/GENSR1802687B
Original scientific paper

PURSUIT FOR EST MICROSATELLITES IN A TETRAPLOID MODEL FROM DE NOVO TRANSCRIPTOME SEQUENCING

Tijana BANJANAC[1], Marijana SKORIĆ[1], Mario BELAMARIĆ[2], Jasmina NESTOROVIĆ ŽIVKOVIĆ[1], Danijela MIŠIĆ[1], Mihailo JELIĆ[2], Slavica DMITROVIĆ[1], Branislav ŠILER[1]

[1]Institute for Biological Research "Siniša Stanković", University of Belgrade, Belgrade, Serbia
[2]Faculty of Biology, University of Belgrade, Belgrade, Serbia

Banjanac T., M. Skorić, M. Belamarić, J. Nestorović Živković, D. Mišić, M. Jelić, S. Dmitrović, B. Šiler (2018): *Pursuit for EST microsatellites in a tetraploid model from de novo transcriptome sequencing*.- Genetika, Vol 50, No.2, 687-703.

Available scientific literature reports very few microsatellite markers derived from tetraploid genomes using *de novo* transcriptome sequencing, mostly because their gain usually represents a major computational challenge due to complicated combinatorics during assembly of sequence reads. Here we present a novel approach for mining polymorphic microsatellite loci from transcriptome data in a tetraploid species with no reference genome available. Pairs of 114 bp long *de novo* sequenced transcriptome reads of *Centaurium erythraea* were merged into short contigs of 170-200 bp each. High accuracy assembly of the pairs of reads was accomplished by a minimum of 14 bp overlap. Sequential bioinformatics operations involved fully free and open-source software and were performed using an average personal computer. Out of the 13 150 candidate contigs harboring SSR motifs obtained in a final output, we randomly chose 16 putative markers for which we designed primers. We tested the effectiveness of the established bioinformatics approach by amplifying them in eight different taxa within the genus *Centaurium* having various ploidy levels (diploids, tetraploids and hexaploids). Nine markers displayed polymorphism and/or transferability among studied taxa. They provided 54 alleles in total, ranging from 2 to 14 alleles per locus. The highest number of alleles was observed in *C. erythraea*, *C. littorale* and a hybridogenic taxon *C. pannonicum.* The developed markers are qualified to be used in genetic population studies on declining natural populations of *Centaurium* species, thus providing valuable

Corresponding author: Tijana Banjanac, Institute for Biological Research "Siniša Stanković", Bul. despota Stefana 142, Belgrade, Serbia; tel/fax: +381112078404; e-mail: tbanjanac@ibiss.bg.ac.rs; orcid.org/0000-0001-9023-9135

information to evolutionary and conservation biologists. The developed cost-effective methodology provides abundant *de novo* assembled short contigs and holds great promise to mine numerous additional EST-SSR-containing markers for possible use in genetics population studies of tetraploid taxa within the genus *Centaurium*.

Key words: microsatellite mining, free and open-source software, tetraploid genome, *Centaurium*, polymorphism

INTRODUCTION

Simple sequence repeats (SSRs) or microsatellites are molecular markers harboring tandem repeated oligonucleotide sequences extensively used in plant population genetics studies (MADESIS *et al.,* 2013). The main preferences of microsatellites in comparison with other types of molecular markers are their co-dominant nature, locus specificity, hypervariability, reproducibility (NYBOM *et al.,* 2014), and ease of detection (SONG *et al.,* 2012). Until the emergence of the NGS technologies and their extensive application, the identification of SSR loci was an expensive, technically and time-consuming mission, and usually returned far fewer loci than were needed to successfully address to major population genetics questions (CASTOE *et al.,* 2015). These problems were even more pronounced when dealing with non-model species (ELLIS and BURKE, 2007). The development of NGS technologies has enabled production of large amounts of sequence data which further allowed rapid and cost-effective development of molecular markers derived from both the genome and transcriptome (EKBLOM and GALINDO, 2011; ZALAPA *et al.,* 2012; PICKETT *et al.,* 2016; WANG *et al.,* 2016). Microsatellites identified from a transcriptome (genic SSRs or EST-SSRs, Expressed Sequenced Tags-Simple Sequence Repeats) are usually derived from partial random sequencing of cDNA libraries. They are more conserved compared to genome-derived microsatellites, having lower applicability in distinguishing closely related genotypes (KALIA *et al.,* 2011). They also harbor less null alleles in comparison to genomic SSRs (ELLIS and BURKE, 2007). On the other hand, they possess higher level of interspecific transferability (e.g. DUFRESNES *et al.,* 2014b; POSTOLACHE *et al.,* 2014) and once developed can be used in comparative genomics across intrageneric species and may even be applicable to higher taxonomical categories (GUPTA *et al.,* 2003; SAHA *et al.,* 2004; VENDRAMIN *et al.,* 2007; FENG *et al.,* 2008; RAVEENDAR *et al.,* 2015).

The major issue in mining SSRs from transcriptome data may be the sequence redundancy that yields multiple sets of markers at the same locus, which can be circumvented by assembling the ESTs into unigenes (CHEN *et al.,* 2015). However, this represents a severe challenge for bioinformatic processing of data especially in tetraploid species and if no reference genome is available (VIJAY *et al.,* 2013; VUKOSAVLJEV *et al.,* 2015). This is due to complex combinatorics which would result in tremendously large amount of data. Besides that, the effective usage of EST-SSRs in polyploid organisms still suffers from issues such as: allele dosage uncertainty, uneven amplification of alleles, detection of null alleles and cloning necessity (DUFRESNE *et al.,* 2014a). Nevertheless, in their recent review, DUFRESNE *et al.,* (2014a) anticipate that rapid NGS development and new computational approaches should dramatically reshape population genetics of polyploids in the near future.

No literature data are available considering the development of SSR markers for the genus *Centaurium* Hill. Common centaury (*Centaurium erythraea* Rafn) is one of the most important pharmacological species from the family Gentianaceae (GRIEVE, 1971) having a long tradition in usage in medicine. It is used for producing plant drug „Centaurii herba“ which is

common ingredient of many commercial formulations (BOTION *et al.*, 2005; NABER, 2013) and is used worldwide as a natural aroma (NEWALL *et al.*, 1996). Genus *Centaurium* comprises about 20 species (MANSION, 2004) and most of them readily hybridize (GUGGISBERG *et al.*, 2006; BANJANAC *et al.*, 2014). Because of that, there are still ambiguities regarding the genus taxonomy and species relations (see MANSION *et al.*, 2005; the Euro+Med PlantBase, 2017; The Plant List, 2017). The development and employment of new molecular markers designed specifically for the genus *Centaurium*, such as genic microsatellites, may be a powerful tool for resolving the taxonomic issues as well as for conservation efforts and sustainable usage of species germplasm.

The main objective of this study was to develop a bioinformatics approach to mine transcriptome-derived microsatellite markers out of short sequences, specifically assembled from raw 114 bp Illumina reads, using tetraploid *C. erythraea* as a model, with minimal financial and computing resources input. Therefore, in the whole process of the identification of microsatellite loci we opted for open source and free software components, applicable on an average performance computer. The principal idea was to evaluate whether our sequential bioinformatics operations have the ability to provide informative, polymorphic and transferable EST-SSR markers applicable to a tetraploid genome, which may be further used in population genetics studies.

MATERIALS AND METHODS

Sequencing and bioinformatics pipeline

Total RNA was extracted according to GASIC *et al.*, (2004) from leaves of a 4 week-old *in vitro* cultivated *Centaurium erythraea* individual, originating from the Luštica Peninsula, Montenegro. After the treatment with DNAse I (Thermo Scientific, Lithuania), reverse transcription was performed using the GeneAmp® Gold RNA PCR Reagent Kit (Applied Byosystems®, UK), with oligo-dT primers, according to the manufacturer`s recommendations. Paired-end sequencing of double-stranded cDNA libraries was conducted on Illumina GAIIx platform. Short read sequences (114 bp each), were recorded in FASTQ format.

The processing of the obtained data was run on a virtual machine with two processor cores and 8 GB of RAM using Linux operating system with a 64-bit kernel. The procedure was divided into six phases which covered steps from raw FASTQ inputting to the construction of primers for PCR amplifications of EST-SSR loci.

In phase I, Trimmomatic software (v. 0.32, BOLGWER *et al.*, 2014), which is especially designed to both clip and trim Illumina paired-end reads, was used (Figure 1 – phase I). The software was set to remove the residual Illumina primer sequences attached to the reads, while low-quality sequences were eliminated as well (reads with a base quality less than 20). In order to find and combine pairs of corresponding forward and reverse reads into single sequences (Figure 1 – phase II) we used a free software Pear (v. 0.9.6, ZHANG *et al.*, 2014). Minimum overlap for assembling sequence pairs was set to 14 bp. The data were converted to FASTA format using the 'sed' command in Linux. Obtained short contigs were input into another free software, GMATo (v. 1.2, WANG *et al.*, 2013), in order to mine sequences with SSR-containing regions (Figure 1 – phase III). The program was set up to identify all sequences with the following microsatellite characteristics: minimum length of k-mer = 3, minimum repetition of k-mers in sequence = 5. Thus, the sequences harboring three to six nucleotide motifs and with at least 5 repeats were finally considered as the potential SSR markers.

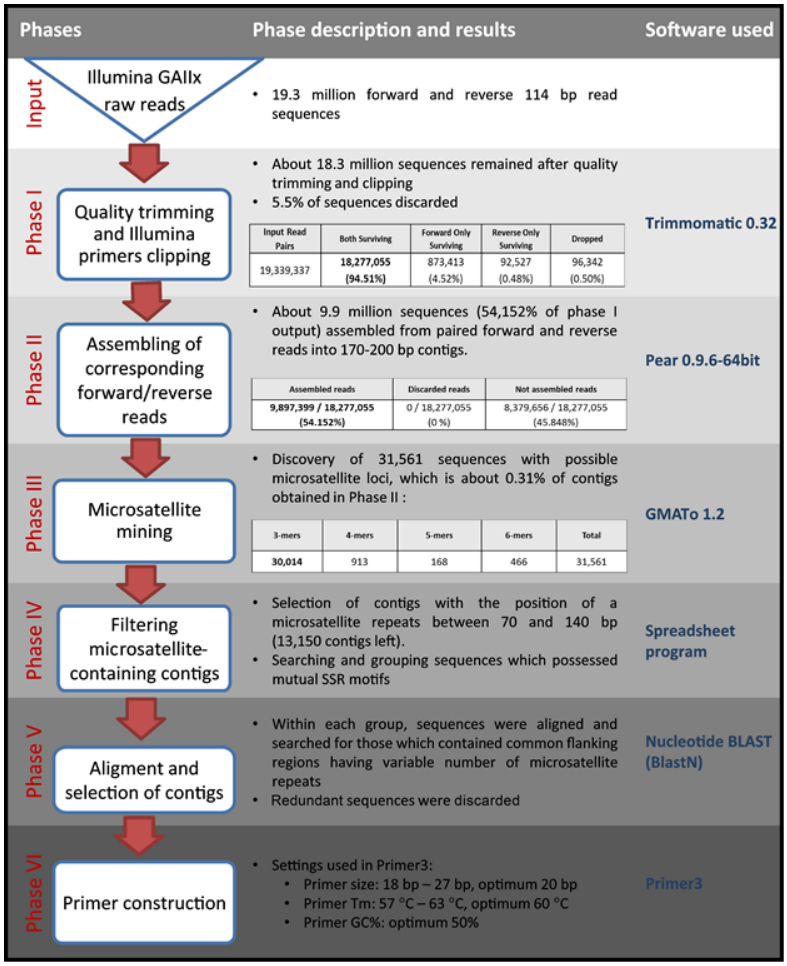


Figure 1. A scheme of the developed bioinformatics pipeline used to effectively obtain polymorphic microsatellite loci from the transcriptome of a tetraploid species *Centaurium erythraea*

Further narrowing the list of possible EST-SSR loci produced by GMATo was accomplished using a spreadsheet software (Figure 1 – phase IV). Our goal was to select all sequences with optimal positioning of microsatellite repeats, i.e. between 70th and 140th bp which would leave sufficient flanking regions for primer design. Among the obtained sequences we randomly selected 16 groups, each of them containing mutual microsatellite motif. BlastN (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to cross-check nucleotide arrays of the sequences within each group and the redundant ones were manually discarded. Sequences with different number of microsatellite repeats, but having highly similar order of nucleotides in flanking regions, were considered potentially polymorphic and were selected for primer design (Figure 1 – phase V). Among the selected sequences, there were two with hexanucleotide SSR motifs, one with pentanucleotide, two with tetranucleotide and eleven with trinucleotide motifs. In the end, software Primer3 (ROZEN and SKALETSKY, 2000) was employed for designing the

primers to amplify the putative polymorphic microsatellite loci (Figure 1 – phase VI). The following parameters were set in Primer 3: primer length range from 18 to 27 bp; PCR product size 100-200 bp; melting temperature between 57°C and 63°C, with 60°C as the optimal annealing temperature; and GC content of 40%-60%, with optimum of 50%.

Evaluation of the EST-SSR markers

Plant material

Seventy plants originating from 14 populations and belonging to eight different taxa of the genus *Centaurium* (seeds deposited at the seed bank of the Institute for Biological Research "Siniša Stanković", Serbia) were used as panel for validation of developed EST-SSR markers' validation, and polymorphism and transferability testing (Table 1). The selected taxa included five *Centaurium* species of different ploidy levels: *C. maritimum* (L.) Fritsch (2x), *C. pulchellum* (Sw.) Druce (4x), *C. tenuiflorum* (Hoffmanns. & Link) Fritsch (2x), *C. erythraea* Rafn (4x), *C. littorale* (Turner) Gilmour with two subspecies: *C. littorale* ssp. *littorale* (4x) and *C. littorale* ssp. *compressum* (4x), and one ex-*Centaurium* species: *Schenkia spicata* (L.) Mansion (=*Centaurium spicatum* (L. Fritsch) (2x). The sample set also contained hexaploid specimens here named *Centaurium pannonicum* (BANJANAC *et al.*, 2017), putatively determined as interspecies hybrids between *C. erythraea* and *C. littorale* ssp. *compressum* using taxonomical keys available (MANSION, 2004) and comparing their morphological characteristics and phytochemical profiles (BANJANAC *et al.*, 2017). The leaf samples were collected from nature, but in some cases, the plants were grown from seeds in a greenhouse.

Table 1. The sample set of 70 individual plants belonging to eight different taxa of the genus Centaurium used in the validation, variability and transferability assays of the potential EST-SSR markers

Populations	Locality	Geographic latitude and longitude	No. of individuals	Estimated ploidy level
C. spicatum	Tivat	42°24‘45.24”N 18°43‘07.41”E	5	2x
C. maritimum	Podgorica	42°27‘25.80”N 19°15‘03.38”E	5	2x
C. pulchellum	Nikšić	42°44‘47.58”N 18°51‘57.39”E	3	4x
C. pulchellum	Palić	46°01‘48.61”N 19°44‘27.96”E	5	4x
C. tenuiflorum	Žanjice	42°23‘45.28”N 18°34‘44.89”E	6	2x
C. tenuiflorum	Sutorina	42°27‘15.18”N 18°29‘53.83”E	6	2x
C. erythraea	Ulcinj	41°53‘30.40”N 19°18‘04.20”E	6	4x
C. erythraea	Beočin	45°10‘34.05”N 19°43‘09.83”E	5	4x
C. erythraea	Bački Vinogradi	46°07‘30.56”N 19°50‘57.43”E	5	4x
C. pannonicum*	Majdan	46°09‘25.50”N 19°36‘31.56”E	4	6x
C. pannonicum*	Palić	46°01‘48.61”N 19°44‘27.96”E	5	6x
C. littorale compressum	Majdan	46°09‘25.50”N 19°36‘31.56”E	5	4x
C. littorale compressum	Palić	46°01‘48.61”N 19°44‘27.96”E	5	4x
C. littorale littorale	Nant	Seeds obtained by *Kew Garden*-France	5	4x

DNA extraction and PCR amplification

Total genomic DNA was isolated from fresh leaf samples using a modified CTAB method (DOYLE and DOYLE, 1990). DNA Concentration and purity of isolates were assessed upon spectrophotometrical absorbance measuring at 260, 280, and 230 nm (Agilent 8453, Agilent Technologies, Germany). Optimization of primers' annealing temperatures and all further PCR amplifications were carried out using Eppendorf Mastercycler nexus gradient thermal cycler (Eppendorf AG, Germany) in a final volume of 25 µl, each reaction containing 100 ng of template DNA, 1 × Taq Buffer ($(NH_4)_2SO_4$), 2.5 mM $MgCl_2$, 1 U Taq DNA polymerase (Thermo Scientific, Lithuania), 200 µM dNTPs (of each dATP, dCTP, dGTP and dTTP), and 0.5 µM each forward or reverse primer (Invitrogen, UK). The PCR reaction program used for amplification of all loci was as follows: 94°C for 3 min; 38 cycles of 94°C for 30 s, 53.0°C - 60.0°C (depending on the primer pair, Table 2) for 30 s, 72°C for 45 s, and 72°C for 10 min as a final extension step.

Table 2. Characteristics of primer pairs for amplification of nine microsatellite loci developed from Centaurium erythraea transcriptome sequences. The expected sizes of fragments (in bp) and marker size ranges are scored across a sample set of 70 individuals belonging to eight taxa of the genus Centaurium. Ta = annealing temperature

Locus	Repeat motif	Primer sequence (5'-3')	GC composition [%]	Expected length [bp]	Observed size range [bp]	Ta [°C]
M4	GAAGAT	F: TGAAATGAAACCCCACCTATG	42.86	150	154-184	58.7
		R: GCATCATGTTGAAAGCGAAG	45.00			
M5	TAC	F: TTGTTGACAGAAGAGAGAGAGCA	43.48	100	106-139	56.5
		R: AGAAGCAAATTCAGACATAAATCAA	28.00			
M7	AAAGA	F: AGGCATAGCCCTTTTTCCAT	45.00	100	95-105	56.1
		R: GACCTTCTTCCCACCTTTCC	55.00			
M8	GTC	F: CAGGACGGATATTATTGTGGTTG	43.48	114	113-125	56.5
		R: CATCTGCGTCAGCCATGT	55.56			
M10	ATT	F: TACCCTGGGACAAAAAGCAT	45.00	185	155-227	56.5
		R: TGGTCATAAATCCTGCCTCTG	47.62			
M12	CGA	F: GACGACGACAGTGAGGATGA	55.00	151	250-292	56.5
		R: TTTTTGATCTGTAGTAGGTCAGAATTT	29.63			
M13	GTT	F: GTCGCTTTTCGTCTCCCAAG	55.00	153	145-172	56.5
		R: TTCTACTGCGTCATGGATAATCA	39.13			
M17	TGT	F: AATTAGAGGGATCACTGAATGC	40.91	158	75, 170	56.5
		R: TGGTTAACAGATGGTACCACAA	40.91			
M18	TGT	F: TGCTCTGGTTTGTTCAAAGG	45.00	173	175-181	58
		R: TCCTTCCTCCTCTTCCTCCT	55.00			

Testing the informativeness, functionality, polymorphism and transferability of the EST-SSR markers

We firstly validated primers developed for amplification of all 16 microsatellite loci by amplifying them in the same individual of *C. erythraea* used for obtaining transcriptome data (originating in Luštica Peninsula, Montenegro). The PCR products were run on 2.5% agarose gels in 1×Tris/borate/EDTA (TBE) buffer at 1.5 V cm^{-1} for 1.5 h, stained with ethidium bromide and visualized by a UV transilluminator (ST4 3026-WL/26M, Vilber Lourmat, France). The 50 bp DNA Ladder (Thermo Scientific, Lithuania) was used for the initial assessment of the amplicons' length.

We then tested transferability and polymorphism of loci by amplifying them in a full sample set of 70 individuals using the Lab-on-a-Chip technology on the Agilent Bioanalyzer 2100 system (DNA 1000 LabChip). Prior to capillary electrophoresis, products of PCR amplification of pairs of microsatellite loci differing in predicted sizes were mixed ("multipooled") along with the DNA fragments of known length, which represented internal standards. Multipooling was performed according to the following scheme:

PCR products of M4 and M7 loci (1 µl each) + 150 bp internal standard (0.3 µl) + 415 bp internal standard (0.3 µl); PCR products of M8 and M10 loci (1 µl each) + 150 bp internal standard (0.3 µl) + 415 bp internal standard (0.3 µl); PCR products of M12 and M13 loci (1 µl each) + 200 bp internal standard (0.3 µl) + 415 bp internal standard (0.3 µl); PCR products of M5 and M18 loci (1 µl each) + 150 bp internal standard (0.3 µl) + 415 bp internal standard (0.3 µl).

We used the commercial DNA fragments as internal standards (NoLimits 150 bp and 200 bp DNA fragment, ThermoFisher Scientific, Bremen, Germany). Another internal standard was obtained by PCR amplification with gene specific primers that produce a fragment of 415 bp. The product of amplification was run on electrophoresis and the amplicon of 415 bp was isolated from the gel using Silica Bead DNA Gel Extraction Kit (Thermoscientific, Bremen, Germany). The concentration was determined fluorimetrically using a Qubit Fluorimeter (Invitrogen®, Carlsbad, CA, USA). All DNA fragments used as internal standards were diluted to a concentration of 50 ng μl^{-1} and as such were used in mixtures' preparation.

Peaks of electropherograms were aligned using Agilent 2100 Expert software (Agilent Technologies, USA), proclaiming the peaks belonging to the internal standards as the lower and upper marker instead of the original ones. This was accomplished by the command "Set Lower/Upper Marker". Due to the difference in size range of the studied loci, the peaks of electropherograms were aligned as follows: 150 bp internal standard was set as the upper marker, while remaining the original lower marker of 15 bp for scoring the alleles of M5, M7 and M8 loci; 150 bp internal standard was set as the lower marker, while 415 bp internal standard was set as the upper marker for the loci M4, M10 and M18; for allele scoring at the locus M13 we retained the original lower marker and set 200 bp as the upper one; and for the locus M12 we selected 200 bp as the lower marker and 415 as the upper marker. Note that the ninth marker, M17, was not analyzed using Lab-on-a-Chip, but instead using 3% agarose electrophoresis, since only two alleles were recorded.

Having the alleles ascertained, we moved on to the determination of genetic distances among studied taxa in order to infer whether the nine randomly selected markers may be used in further population genetic and taxonomical studies. This was performed using R package

polysat, having in mind the difference in ploidy levels among the taxa. We calculated Bruvo distances and visualized them on a PCoA diagram.

RESULTS AND DISCUSSION

Bioinformatics pipeline to discover EST-SSR markers from transcriptome data in a tetraploid species

The basic challenge of this work refers to the development of sequential operations on an average performance personal computer using open source or free software components in order to mine EST-SSR markers within the transcriptome of a tetraploid model. The Illumina sequencing (i.e., sequencing by synthesis) is one of the NGS technologies quite extensively used for microsatellite mining in plants (MUDALKAR *et al.*, 2013; THAMMINA *et al.*, 2014; ZHOU *et al.*, 2016; HODEL *et al.*, 2016). Since the price of Illumina sequencing is in a steady decline considering the cost per Mb (LIU *et al.*, 2012; ZALAPA *et al.*, 2012), this technique represents an obvious choice for researches dealing with understudied species, such as, for instance, those belonging to the genus *Centaurium*. However, random or shotgun sequencing within cDNA libraries usually generates a high ratio of redundant ESTs (VARSHNEY *et al.*, 2004). Regular assembly of a redundant EST dataset, such as short 114 bp reads obtained by Illumina GAIIx platform, might represent a substantial bioinformatics challenge for researches with limited computational resources, expertise and funding (ZALAPA *et al.*, 2012). However, it is possible to first scan a redundant EST library for the sequences that contain microsatellite repeats and then use this significantly smaller dataset in order to identify non-redundant sequences harboring EST-SSRs (KOTA *et al.*, 2001; KANTETY *et al.*, 2002; THIEL *et al.*, 2003; VARSHNEY *et al.*, 2004, 2005). The latter approach proved to be the method of choice for our tetraploid plant model.

In our study, about 19.3 million short read sequences were obtained with Illumina GAIIx sequencing of the *Centaurium erythraea* transcriptome, representing a dataset of about 2205 million base pairs. Two paired-end data files were recorded in FASTQ format, weighting in about of 11.2 GB of data combined. As the initial data processing step, it was necessary to clip Illumina specific primers that were left after sequencing and to filter out low quality read nucleotide arrays. Trimmomatic, a software written in Java (program language widely used in bioinformatics, HOLLAND *et al.*, 2008), was used for these processes. As a result of this initial step, in throwing out about one million sequences (Figure 1 – phase I), and two sets of FASTQ files were obtained, one with forward and one with reverse read sequences.

As stated above, paired-end sequencing of double-stranded cDNA libraries was obtained from a tetraploid species. In such case, assembling reads into unigenes would require exceptionally large amount of bioinformatics resources as well as construction of specific algorithms able to successfully process only a minor portion of reads (MALKOV and SIMONOVIĆ, 2011) to obtain contigs which, however, would be of a rather questionable reliability (MALKOV pers. comm.). Therefore, in the presented approach, having no reference genome as a template, we decided to *de novo* create short contigs out of pairs of forward and reverse reads, instead of assembling unigenes. In several studies, it has been reported that microsatellite mining should not necessarily include unigene assembly (CASTOE *et al.*, 2015; ZALAPA *et al.*, 2012; VUKOSAVLJEV *et al.*, 2015). To have this accomplished, the Pear program was used as a fast, memory-efficient and highly accurate pair-end read merger (ZHANG *et al.*, 2014). The program is fully parallelized and can run with as low as just a few kilobytes of memory, and it is distributed under the Creative Commons license (www.creativecommons.org/licenses), running on the

command-line under Linux and UNIX based operating systems. While the default settings for overlapping value is usually 10 nucleotides (ZHANG *et al.,* 2014), we set this value to 14 nucleotides in order to achieve higher accuracy of assembled sequences (MALKOV, pers. comm.). The result of this operation was a merge of over 54% of 18.3 million sequences and a batch of sequences ranging from 170 to 200 bp in length with an average length of 185.2 bp. (Fig 1 - phase II). However, there is a possibility that such short length of the resulted sequences would leave fewer opportunities for multipooling many loci differing in size range. In this study, though, we show that at least pairs of loci might be multipooled without the employment of fluorescently labeled primers.

During the first two phases of the constructed pipeline, the data were filtered and adjusted to meet the requirements of the third, microsatellite-mining phase, which may be performed with numerous alternative software solutions (GROVER *et al.,* 2011; WANG *et al.,* 2013). We used a common algorithm found in previously developed software pipelines for microsatellite mining and picked out the GMATo software (WANG *et al.,* 2013), which operates with a memory footprint small enough to run on an average personal computer with Windows, Linux and MAC OSs. Since GMATo software requires input files in FASTA format, the quality description of the sequenced material (50% data of FASTQ format) was discarded prior processing. Out of the 9,897,399 contigs, the mining process has discovered 31,561 contigs containing microsatellite motifs (0.31% of the contigs obtained in the phase II). Nearly 95% of these contigs harbored trinucleotide motifs, while 4-mers, 5-mers and 6-mers altogether were represented by less than 5% (Fig. 1 - phase III).

We further found 13 150 contigs containing microsatellite motifs nested in the optimal position (between 70^{th} and 140^{th} bp). Grouping contigs based on mutual microsatellite motifs and consequential *in silico* search for polymorphisms enabled us to find informative and polymorphic microsatellite loci prior to electrophoretic testing.

Microsatellite variation within the sample set

During the validation process, nine loci were successfully amplified in *C. erythraea* using newly developed primers, and high-quality and reproducible bands were obtained upon agarose electrophoresis (Table 2). PCR products were not obtained with primer pairs designed for the amplification of five loci: M1, M2, M3, M6 and M9. This is concordant with some previous studies reporting 60-90% successful amplifications for genic SSR loci, though obtained by different and more expensive strategies (KOTA *et al.,* 2001; THIEL *et al.,* 2003; NICOT *et al.,* 2004; SAHA *et al.,* 2004; CHEN *et al.,* 2015). Additionally, two primer pairs, designed for the amplification of markers M15 and M16, produced multiple amplification products. Unsuccessful amplification of EST-SSR loci may be due to unrecognized intron splice sites which may disrupt priming sequences or, alternatively, the presence of large introns that fall between the primers, thus resulting in large PCR products or, in extreme cases, to the complete PCR amplification failure (ELLIS and BURKE, 2007). Out of the nine successfully amplified loci, the PCR products of eight were of the expected lengths while primer pair designed for the marker M12 gave considerably longer products (Table 2). This phenomenon is reported to be more common for amplicons obtained by genic SSR primers than those obtained by primers designed for genomic SSRs (KOTA *et al.,* 2001; THIEL *et al.,* 2003; NICOT *et al.,* 2004; YU *et al.,* 2004), and is probably due to the presence of introns and/or insertions/deletions (indels) in the corresponding genomic sequence (SAHA *et al.,* 2004).

The process of microsatellite markers development should preferably include the assessment of their polymorphism and transferability in order to reduce costs and increase feasibility of population genetics studies (REID *et al.*, 2012). Although EST-SSRs may be highly polymorphic (e.g., ALEKSIĆ *et al.*, 2009), they are generally considered less polymorphic compared to genomic SSRs because they originate from a more conserved parts of a genome (RUSSELL *et al.*, 2004; CHABANE *et al.*, 2005; SULLIVAN *et al.*, 2013; POSTOLACHE *et al.*, 2014; ZHOU *et al.*, 2016); on the other hand this enables higher transferability of EST-SSRs compared to genomic SSRs, making them more suitable for studies dealing with species relations (GUPTA *et al.*, 2003; VENDRAMIN *et al.*, 2007; FENG *et al.*, 2008; KUMAR *et al.*, 2013). Therefore, the next step in our study was to perform a detailed evaluation of the applicability of the *C. erythraea*-derived EST-SSR markers to other species of the genus and to assess their polymorphism.

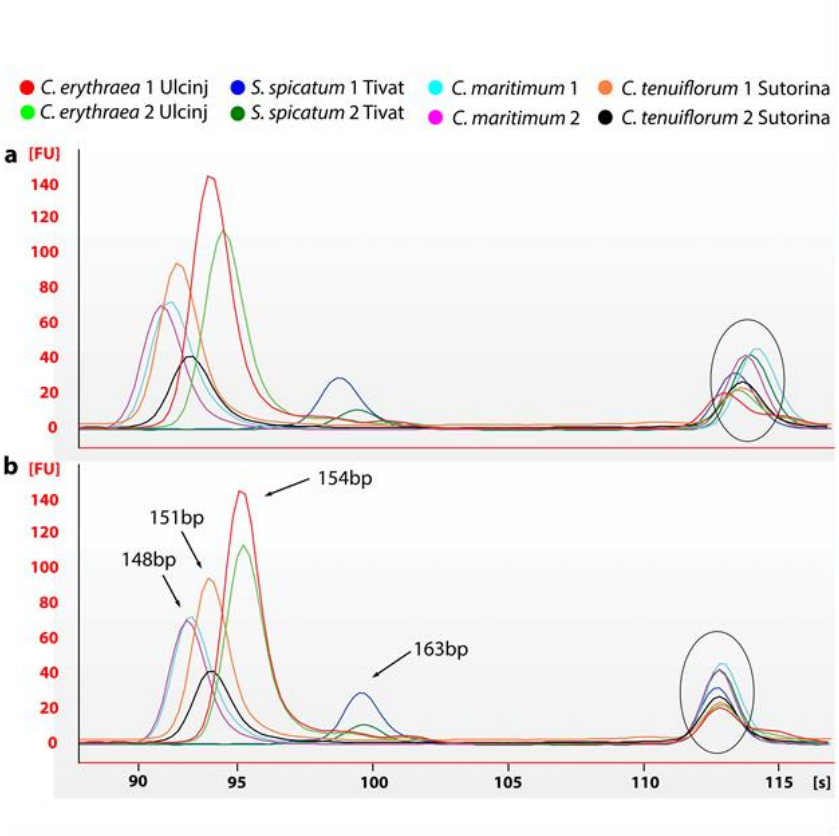


Figure 2. Overlapped electropherograms of a lokus M13 amplified in selected Centaurium species and visualized by the Agilent Bioanalyzer 2100 Expert software a) using the default settings with the original upper and lower marker and b) using the internal size standards for peak alignment. For the interpretation of colors in the pictures, the reader is referred to the electronic version of the article

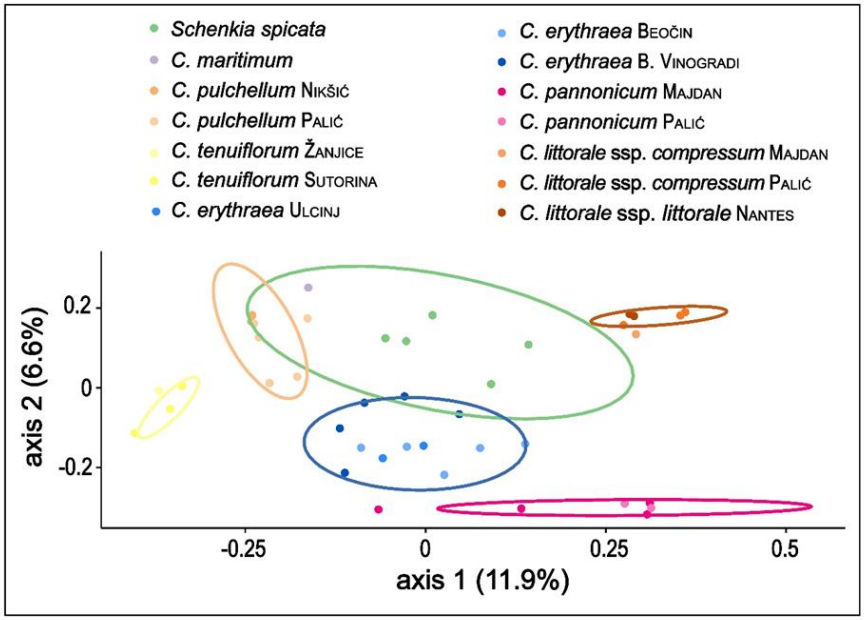
The following putative microsatellite markers: M4, M5, M7, M8, M10, M12, M13, M17 and M18 were amplified in 70 individuals in order to assess their polymorphism and transferability (Supplementary table 1). We opted to score microsatellite allele lengths using the Agilent 2100 Bioanalyzer employing Lab-on-a-chip technology, which is based on capillary electrophoresis. The device is equipped with a fluorescent detection system that offers high detection sensitivity, while the fragment size is estimated by comparison with high-precision standards (PANARO *et al.*, 2000). The accuracy of the fragment sizing of the instrument using the DNA 1000 Assay in the range between 25 and 500 bp is 5% as reported by the manufacturer. We circumvented this issue by the introduction of internal size standards (Figure 2), which do not interrupt the marker size range, but have their own peaks in the vicinity of the expected alleles' peaks. In such way, we achieved a-base-pair-precise sizing, suitable for both across-sample and across-chip comparison of the allele lengths (Figure 2). For each locus, we observed shifts in fragment sizes corresponding to the length of a particular microsatellite motif. We have also been able to multipool PCR products of pairs of loci with different size ranges that do not overlap with the internal size standards. Thus, a cost-effective assay was established for usage in analyses which include small sample sets (less than 200 runs): the initial finance input per sample does not involve purchase of fluorescently labelled primers nor costs for capillary electrophoresis of fluorescently labelled products. There are only few reports in which the same methodology has been used for the scoring of microsatellite alleles (FRANCISCO-CANDEIRA *et al.*, 2007; VARELA *et al.*, 2007; MUZZALUPO *et al.*, 2010).

For the nine successfully amplified loci, amplification products were reproducible and all loci were polymorphic. In overall sample set, we recorded 54 alleles in total (Supplementary table 1). The number of alleles per locus ranged from 2 for the locus M17 to 14 for the locus M10 with the mean number of 6 alleles per locus. When considering the number of alleles per population across all loci, we observed higher values in two populations of *C. erythraea* and both populations of *C. pannonicum* (Supplementary table 1) than in other studied species. This might be explained by the fact that the markers were developed with the idea to be variable as much as possible in the tetraploid species *C. erythraea*, which is also reported to be one of the parental species of the hybridogenic taxon *C. pannonicum*.

Marker M4 was not amplified in one *C. tenuiflorum* population (Žanjice), while in the second population of the same species (Sutorina), only one allele was observed at this locus (Supplementary table 1). Moreover, markers M7 and M17 failed to amplify in species *C. pulchellum* and *C. tenuiflorum*, same as markers M7 and M18 in *C. maritimum*. EST-SSR markers which do not amplify in some species, while do amplify in others, may be exceptionally informative in species determination (e.g. SAKAGUCHI and ITO, 2014). The same holds for private population/species alleles, which can be observed in Supplementary table 1. However, well known issues in allele scoring in polyploids, such as allelic ratio at a particular locus and incidence of null alleles, remained unresolved. Thus, for tetraploid species *C. erythraea*, *C. littorale* and *C. pulchellum*, as well as for hexaploid hybridogenic taxon *C. pannonicum*, we could not determine the exact genotypes. Therefore, *polysat* package was used for the calculation of the Bruvo distances (BRUVO *et al.*, 2004) between individuals, which takes into account mutation processes in the estimation of allelic frequencies and permit comparison of individuals with different ploidy levels. Although the set of the nine randomly developed markers represents only a small part of the minable EST-SSRs (the mining process has discovered 31,561 contigs containing microsatellite motifs), it showed a satisfying power to resolve each of the seven

studied species using the Bruvo distances, which is represented in a two-dimensional PCoA plot (Supplementary figure 1).

Nine developed markers represent only a small part of the minable EST-SSRs: the mining process has discovered 31,561 contigs harboring trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide motifs. Although being a common class of microsatellites (YUE *et al.,* 2014) we did not mine for dinucleotide motifs due to possible difficulties in scoring different alleles using presented methodology.



Supplementary figure 1. PCoA scatterplot of 70 individuals belonging to 14 populations of eight different taxa of the genus Centaurium. Ellipses are constructed according to the 95% confidence interval. Note that C. maritimum (a gray dot) is represented with only one genotype with no ellipse supplied, owing lack of polymorphism among the five individuals. C. littorale is represented with a joint ellipse, regardless the subspecies

CONCLUSIONS

The significance of the presented work is that we have developed a bioinformatics pipeline that spares time and costs to obtain polymorphic and transferable microsatellite markers for a tetraploid plant species. We used transcriptome sequences of *C. erythraea*, made short contigs of 170-200 bp, mined EST-SSRs and, for all data processing, employed common hardware and software products available in ordinary labs with limited funding. The successfulness of this approach was supported by observed high transferability and polymorphic nature of the developed EST-SSR markers. Nine out of the 16 randomly selected potential EST-

SSR markers showed considerable polymorphism across the sample set tested. To the best of our knowledge, this is the first successful attempt to create and validate microsatellite markers for any of the *Centaurium* species. Future studies on tetraploids aiming at resolving issues regarding the determination of the allele contribution in peaks (allelic dosage) are required, and that would enable the assessment of allele frequencies necessary for comprehensive studies of genetic relations within the genus *Centaurium.*

ACKNOWLEDGEMENTS

This research was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia (grant No. 173024). The authors would like to thank Dr. Saša Malkov for providing precious bioinformatics advices and to Dr. Valentina Đorđević and Dr. Dragica Radojković for providing access to Agilent Bioanalyzer. The authors would like to thank Dr. Ana Simonović for professional help. The visual solution of the PCoA scatterplot is credited to Dr. Milan Dragićević.

Received, October 10th, 2017
Accepted, April 18th, 2018

REFERENCES

ALEKSIĆ, J.M., S., SCHUELER, M., MENGL, T., GEBUREK (2009): EST-SSRS developed for other *Picea* species amplify in *Picea omorika* and reveal high genetic variation in two natural populations. Belg. J. Bot., *142*: 89-95.

BANJANAC, T., M., DRAGIČEVIĆ, B., ŠILER, U., GAŠIĆ, B., BOHANEC, NESTOROVIĆ J., ŽIVKOVIĆ, S., TRIFUNOVIĆ, D., MIŠIĆ (2017): Chemodiversity of two closely related tetraploid *Centaurium* species and their hexaploid hybrid: Metabolomic search for high-resolution taxonomic classifiers. Phytochemistry, *140*: 27-44.

BANJANAC, T., B., ŠILER, M., SKORIĆ, N., GHALAWENJI, M., MILUTINOVIĆ, D., BOŽIĆ, D., MIŠIĆ (2014): Interspecific in vitro hybridization in genus *Centaurium* and identification of hybrids via flow cytometry, RAPD, and secondary metabolite profiles. Turk. J. Bot., *38*: 68–79.

BOLGER, A.M., M., LOHSE, B., USADEL (2014): Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, *30*(15): 2114–2120.

BOTION, L.M., A.V.M., FERREIRA, S.F., CÔRTES, V.S., LEMOS, F.C., BRAGA (2005): Effects of the Brazilian phytopharmaceutical product Ierobina® on lipid metabolism and intestinal tonus. J. Ethnopharmacol., *102*(2): 137–142.

BRUVO, R., N., MICHIELS, T., D'SOUZA, H., SCHULENBURG (2004): A Simple Method for the Calculation of Microsatellite Genotype Distances Irrespective of Ploidy Level. Mol. Ecol., *13*(7): 2101–2106.

CASTOE, T.A., A.W., POOLE, A.P.J., DE KONING, K.L., JONES, D.F., TOMBACK, S.J., OYLER-MCCANCE, J.A., FIKE, S.L., LANCE, J.W., STREICHER, E.N., SMITH, D.D., POLLOCK (2015): Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. PLOS ONE, *10*(8): e0136465.

CHABANE, K., G.A., ABLETT, G.M., CORDEIRO, J., VALKOUN, R.J., HENRY (2005): EST versus genomic derived microsatellite markers for genotyping wild and cultivated barley. Gen. Res. Crop Evol., *52*(7): 903–909.

CHEN, H., L., LIU, L., WANG, S., WANG, P., SOMTA, X., CHENG (2015): Development and validation of EST-SSR markers from the transcriptome of adzuki bean (*Vigna angularis*). PLOS ONE, *10*(7): e0131939.

DOYLE, J., J., DOYLE (1990): Isolation of plant DNA from leaf tissue. Focus, *12*: 13–15.

DUFRESNE, F., M., STIFT, R., VERGILINO, B.K., MABLE (2014a): Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. Mol. Ecol., *23*(1): 40–69.

DUFRESNES, C., A., BRELSFORD, P., BÉZIERS, N., PERRIN (2014b): Stronger transferability but lower variability in transcriptomic- than in anonymous microsatellites: evidence from Hylid frogs. Mol. Ecol. Resour., *14*(4): 716–725.

EKBLOM, R., J., GALINDO (2011): Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, *107*(1): 1–15.

ELLIS, J.R., J.M., BURKE (2007): EST-SSRs as a resource for population genetic analyses. Heredity, *99*:125–132.

FENG, S.P., W.G., LI, H.S., HUANG, J.Y., WANG, Y.T., WU (2008): Development, characterization and cross-species/genera transferability of EST-SSR markers for rubber tree (*Hevea brasiliensis*). Mol. Breed., *23*(1): 85–97.

FRANCISCO-CANDEIRA, M., A., GONZÁLEZ-TIZÓN, M.A., VARELA, A., MARTÍNEZ-LAGE (2007): Development of microsatellite markers in the razor clam *Solen marginatus* (Bivalvia: Solenidae). J. Mar. Biol. Assoc. UK, *87*(4): 977–978.

GASIC, K., A., HERNANDEZ, S.S., KORBAN (2004): RNA Extraction from Different Apple Tissues Rich in Polyphenols and Polysaccharides for cDNA Library Construction. Plant. Mol. Biol. Rep., *22*(4): 437–438.

GRIEVE, M. (1971): A modern herbal. Dover Publications, New York.

GROVER, A., V., AISHWARYA, P.C., SHARMA (2011): Searching microsatellites in DNA sequences: approaches used and tools developed. Physiol. Mol. Biol. Plants, *18*(1): 11–19.

GUGGISBERG, A., F., BRETAGNOLLE, G., MANSION (2006): Allopolyploid origin of the mediterranean endemic, *Centaurium bianoris* (Gentianaceae), inferred by molecular markers. Syst. Bot., *31*(2): 368–379.

GUPTA, P.K., S., RUSTGI, S., SHARMA, R., SINGH, N., KUMAR, H.S., BALYAN (2003): Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol. Genet. Genomics, *270*(4): 315–323.

HODEL, R., C., SEGOVIA-SALCEDO, J., LANDIS, A., CROWL, M., SUN, X., LIU, M., GITZENDANNER, N., DOUGLAS, C., GERMAIN-AUBREY, S., CHEN, D., SOLTIS, P., SOLTIS (2016): The Report of My Death Was an Exaggeration: A Review for Researchers Using Microsatellites in the 21st century1. Appl. Plant. Sci., *4*(6): 1600025.

HOLLAND, R.C.G., T.A., DOWN, M., POCOCK, A., PRLIĆ, D., HUEN, K., JAMES, S., FOISY, A., DRÄGER, A., YATES, M., HEUER, M.J., SCHREIBER (2008): BioJava: an open-source framework for bioinformatics. Bioinformatics, *24*(18): 2096–2097.

KALIA, R.K., M.K., RAI, S., KALIA, R., SINGH, A.K., DHAWAN (2011): Microsatellite markers: an overview of the recent progress in plants. Euphytica, *177*(3): 309–334.

KANTETY, R.V., M.L., ROTA, D.E., MATTHEWS, M.E., SORRELLS (2002): Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol. Biol., *48*(5-6): 501–510.

KOTA, R., R.K., VARSHNEY, T., THIEL, K.J., DEHMER, A., GRANER (2001): Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas, *135*(2-3): 145–151.

KUMAR, S., S., RAI, D.K., MAURYA, P.L., KASHYAP, A.K., SRIVASTAVA, M., ANANDARAJ (2013): Cross-species transferability of microsatellite markers from *Fusarium oxysporum* for the assessment of genetic diversity in *Fusarium udum*. Phytoparasitica, *41*(5): 615–622.

LIU, L., Y., LI, S., LI, N., HU, Y., HE, R., PONG, D., LIN, L., LU, M. LAW (2012): Comparison of next-generation sequencing systems. J. Biomed. Biotech,, 1–11.

MADESIS, P., I., GANOPOULOS, A., TSAFTARIS (2013): Microsatellites: Evolution and contribution. In: Microsatellites: Methods and Protocols; Methods in Molecular Biology, Volume *1006*. Edited by Kantartzi S.K.: Springer Science and Business Media, LLC, pp 1-13.

MALKOV, S., A., SIMONOVIĆ (2011): Shotgun assembly of *Centaurium erythraea* transcriptome. 19th Symposium of the Serbian Plant Physiology Society. Book of abstracts, 16.

MANSION, G. (2004): A new classification of the polyphyletic genus *Centaurium* Hill (Chironiinae, Gentianaceae): description of the new world endemic *Zeltnera*, and reinstatement of *Gyrandra* Griseb. and *Schenkia* Griseb. Taxon, *53*(3): 719–740.

MANSION, G., L., ZELTNER, F., BRETAGNOLLE (2005): Phylogenetic patterns and polyploid evolution within the mediterranean genus *Centaurium* (Gentianaceae-Chironieae). Taxon, *54*: 931–950.

MUDALKAR, S., R., GOLLA, S., GHATTY, A.R., REDDY (2013): De novo transcriptome analysis of an imminent biofuel crop, *Camelina sativa* L. using Illumina GAIIX sequencing platform and identification of SSR markers. Plant. Mol. Biol., *84*(1-2): 159–171.

MUZZALUPO, I., A., CHIAPPETTA, C., BENINCASA, E., PERRI (2010): Intra-cultivar variability of three major olive cultivars grown in different areas of central-southern Italy and studied using microsatellite markers. Sci. Hortic., *126*(3): 324–329.

NABER, K.G. (2013): Efficacy and safety of the phytotherapeutic drug Canephron® N in prevention and treatment of urogenital and gestational disease: review of clinical experience in Eastern Europe and Central Asia. Res. Rep. Urol., *5*:39–46.

NEWALL, C.A., L.A., ANDERSON, J.D., PHILLIPSON (1996): Herbal medicines. A guide for health-care professionals. The Pharmaceutical Press, London.

NICOT, N., V., CHIQUET, B., GANDON, L., AMILHAT, F., LEGEAI, P., LEROY, M., BERNARD, P., SOURDILLE (2004): Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). TAG, *109*(4): 800–805.

NYBOM, H., K., WEISING, B., ROTTER (2014): DNA fingerprinting in botany: past, present, future. Investig. Genet., *5*:1.

PANARO, N.J., P.K., YUEN, T., SAKAZUME, P., FORTINA, L.J., KRICKA, P., WILDING (2000): Evaluation of DNA fragment sizing and quantification by the Agilent 2100 bioanalyzer. Clin. Chem., *46*(11): 1851–1853.

PICKETT, B.D., S.M., KARLINSEY, C.E., PENROD, M.J., CORMIER, M.T.W., EBBERT, D.K., SHIOZAWA, C.J., WHIPPLE, P.G., RIDGE (2016): SA-SSR: a suffix array-based algorithm for exhaustive and efficient SSR discovery in large genetic sequences. Bioinformatics, *32*(17): 2707-2709.

POSTOLACHE, D., C., LEONARDUZZI, A., PIOTTI, I., SPANU, A., ROIG, B., FADY, A., ROSCHANSKI, S., LIEPELT, G.G., VENDRAMIN (2014): Transcriptome versus genomic microsatellite markers: Highly informative multiplexes for genotyping *Abies alba* Mill. and congeneric species. Plant. Mol. Biol. Rep., *32*(3): 750–760.

RAVEENDAR, S., G-A., LEE, Y-A., JEON, Y.J., LEE, J-R., LEE, G-T., CHO, J-H., CHO, J-H., PARK, K-H., MA, J-W., CHUNG (2015): Cross-amplification of *Vicia sativa* subsp. *sativa* microsatellites across 22 other *Vicia* species. Molecules, *20*(1): 1543–1550.

REID, K., T.B., HOAREAU, P., BLOOMER (2012): High-throughput microsatellite marker development in two sparid species and verification of their transferability in the family Sparidae. Mol. Ecol. Resour., *12*(4): 740–752.

ROZEN, S., H., SKALETSKY (2000): Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol., *132*:365–386.

RUSSELL, J., A., BOOTH, J., FULLER, B., HARROWER, P., HEDLEY, G., MACHRAY, W., POWELL (2004): A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome, *47*(2): 389–398.

SAHA, M.C., M.A.R., MIAN, I., EUJAYL, J.C., ZWONITZER, L., WANG, G.D., MAY (2004): Tall fescue EST-SSR markers with transferability across several grass species. TAG, *109*(4): 783–791.

SAKAGUCHI, S., M., ITO (2014): Development and characterization of EST-SSR markers for the *Solidago virgaurea* complex (Asteraceae) in the Japanese archipelago. Appl. Plant. Sci., *2*(7): 1400035.

SONG, Y-P., X-B., JIANG, M., ZHANG, Z-L., WANG, W-H., BO, X-M., AN, D-Q., ZHANG, Z-Y., ZHANG (2012): Differences of EST-SSR and genomic-SSR markers in assessing genetic diversity in poplar. Forestry Studies in China, *14*(1): 1-7.

SULLIVAN, A.R., J.F., LIND, T.S., MCCLEARY, J., ROMERO-SEVERSON, O., GAILING (2013): Development and characterization of genomic and gene-based microsatellite markers in North American red oak species. Plant. Mol. Biol. Rep., *31*(1): 231–239.

THAMMINA, C.S., R.T., OLSEN, M., MALAPI-WIGHT, J.A., CROUCH, M.R., POOLER (2014): Development of polymorphic genic-SSR markers by cDNA library sequencing in boxwood, *Buxus* spp. (Buxaceae). Appl. Plant. Sci., *2*(12): 1400095.

THE EURO+MED PLANTBASE (2017): Http://Ww2.Bgbm.Org/Europlusmed/Ptaxondetail.Asp?Namecache=Centaurium &Ptreffk =7200000. Accessed 27 March 2017.

THE PLANT LIST (2017): http://www.theplantlist.org/tpl1.1/search?q=centaurium. Accessed 27 March 2017.

THIEL, T., W., MICHALEK, R.K., VARSHNEY, A., GRANER (2003): Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). TAG, *106*(3): 411–422.

VARELA, M.A., A., GONZÁLEZ-TIZÓN, L., MARIÑAS, A., MARTÍNEZ-LAGE (2007): Genetic divergence detected by ISSR markers and characterization of microsatellite regions in *Mytilus mussels*. Biochem. Genet., *45*(7-8): 565–578.

VARSHNEY, R.K., A., GRANER, M.E., SORRELLS (2005): Genic microsatellite markers in plants: features and applications. Trends Biotechnol., *23*(1): 48–55.

VARSHNEY, R.K., H., ZHANG, E., POTOKINA, N., STEIN, P., LANGRIDGE, A., GRANER (2004): A simple hybridization-based strategy for the generation of non-redundant EST collections - a case study in barley (*Hordeum vulgare* L.). Plant. Sci., *167*(3): 629–634.

VENDRAMIN, E., M.T., DETTORI, J., GIOVINAZZI, S., MICALI, R., QUARTA, I., VERDE (2007): A set of EST-SSRs isolated from peach fruit transcriptome and their transportability across *Prunus* species. Mol. Ecol. Notes, *7*(2): 307–310.

VIJAY, N., J.W., POELSTRA, A., KÜNSTNER, J.B.W. WOLF (2013): Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. Mol. Ecol., *22*(3): 620–634.

VUKOSAVLJEV, M., G.D., ESSELINK, W.P.C., VAN'T WESTENDE, P., COX, R.G.F., VISSER, P., ARENS, M.J.M., SMULDERS (2015): Efficient development of highly polymorphic microsatellite markers based on polymorphic repeats in transcriptome sequences of multiple individuals. Mol. Ecol. Resour., *15*(1): 17–27.

WANG, P., L., YANG, E., ZHANG, Z., QIN, H., WANG, Y., LIAO, X., WANG, L., GAO (2016): Characterization and development of EST-SSR markers from a cold-stressed transcriptome of centipedegrass by Illumina paired-end sequencing. Plant. Mol. Biol. Rep., *35*(2): 215-223.

WANG, X., P., LU, Z., LUO (2013): GMATo: A novel tool for the identification and analysis of microsatellites in large genomes. Bioinformation, *9*(10): 541–544.

YU, J-K., T.M., DAKE, S., SINGH, D., BENSCHER, W., LI, B., GILL, M.E., SORRELLS (2004): Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. Genome, *47*(5): 805–818.

YUE, X-Y., G-Q., LIU, Y., ZONG, Y-W., TENG, D-Y., CAI (2014): Development of genic SSR markers from transcriptome sequencing of pear buds. J. Zhejiang. Univ. Sci. B., *15*(4):303-312.

ZALAPA, J.E., H., CUEVAS, H., ZHU, S., STEFFAN, D., SENALIK, E., ZELDIN, B., MCCOWN, R., HARBUT, P., SIMON (2012): Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am. J. Bot., *99*(2): 193–208.

ZHANG, J., K., KOBERT, T., FLOURI, A., STAMATAKIS (2014): PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, *30*(5): 614–620.

ZHOU, Q., D., LUO, L., MA, W., XIE, Y., WANG, Y., WANG, Z., LIU (2016): Development and cross-species transferability of EST-SSR markers in Siberian wildrye (*Elymus sibiricus* L.) using Illumina sequencing. Sci. Rep., *6*:20549.

POTRAGA ZA EST-SSR MARKERIMA NA OSNOVU PODATAKA *DE NOVO* SEKVENCIRANJA TRANSKRIPTOMA TETRAPLOIDNE MODEL VRSTE

Tijana BANJANAC[1], Marijana SKORIĆ[1], Mario BELAMARIĆ[2], Jasmina NESTOROVIĆ ŽIVKOVIĆ[1], Danijela MIŠIĆ[1], Mihailo JELIĆ[2], Slavica DMITROVIĆ[1], Branislav ŠILER[1]

[1]Institut za biološka istraživanja "Siniša Stanković", Univerzitet u Beogradu, Beograd
[2]Biološki fakultet, Univerzitet u Beogradu, Beograd

Izvod

U modernoj naučnoj literaturi postoji svega nekoliko naučnih publikacija koje se bave razvijanjem mikrosatelitskih markera na osnovu transkriptoma ili genoma tetraploidnih vrsta. Dobijanje mikrosatelitskih markera obično predstavlja poseban računarski izazov usled komplikovane i obimne kombinatorike tokom slaganja (eng.: *assembly*) sekvenciranog materijala u kontige (duže nizove); izazov koji se još dodatno uvećava kada su u pitanju poliploidne vrste. U okviru ovog istraživanja predstavljen je inovativan bioinformatički pristup pronalaženju polimorfnih mikrosatelitskih lokusa u okviru transkriptoma tetraploidne vrste za koju ne postoji referentni genom. Osobenost primenjenog pristupa se ogleda u kreiranju kratkih kontiga, dužine od 170 do 200 baznih parova, na osnovu takođe kratkih (114 bp) uparenih sekvenci (tzv. „ridova"; eng.: *read*) *de novo* sekvenciranog transkriptoma vrste *Centaurium erythraea*. Visoka tačnost uparivanja kratkih sekvenci je postignuta minimalnim preklapanjem od 14 baznih parova. Bioinformatički pristup obradi podataka je podeljen u šest faza za koje su upotrebljeni isključivo besplatni softveri otvorenog koda koji su, takođe, omogućavali i celokupnu obradu podataka na prosečnom personalnom računaru. Kao krajnji rezultat primenjenog pristupa dobijen je set od 13 150 kratkih kontiga koje sadrže mikrosatelitske ponovke od kojih je, po principu slučajnosti, odabrano 16 neredundantnih koje su poslužile za konstrukciju prajmera i potvrdu samog pristupa. Za 9 od 16 konstruisanih lokusa pokazana je polimorfnost i/ili transferabilnost u okviru osam različitih taksona roda *Centaurium*. Ukupno su zabeležena 54 različita alela, dok je broj alela po lokusu varirao u rasponu od 2 do 14. Najveći broj alela je zabeležen kod jedinki *C. erythraea*, *C. littorale* i kod hibridogenog taksona *C. pannonicum*. Dobijeni EST-SSR markeri se mogu iskoristiti u populaciono-genetičkim istraživanjima prirodnih populacija roda *Centaurium* te tako pružiti i dragocene informacije konzervacionim i evolucionim biolozima. Primenjena metodologija je obezbedila veoma veliku bazu podatak *de novo* asemblovanih i odabranih kontiga koja može biti osnova za konstruisanje još velikog broja EST-SSR markera za primenu u obimnijim populaciono-genetičkim istraživanjima tetraploidnih taksona u okviru roda *Centaurium*.

Primljeno 11.X.2017.
Odobreno 18. IV. 2018.