Using Bioinformatics Tools for Identification and Characterization of Transcriptome Derived EST-SSRs in Silver Fir *(Abies alba* Mill.*)*

Dragoş POSTOLACHE*

National Institute for Research and Development in Forestry (INCDS) "Marin Drăcea",
str. Horea 65, 400275 Cluj-Napoca, Romania
***)**Corresponding author, e-mail: dragospostolache@yahoo.com

BulletinUASVM Horticulture 74(1) / 2017
Print ISSN 1843-5254, Electronic ISSN 1843-5394
DOI:10.15835/buasvmcn-hort:12625

Abstract

Bioinformatics tools have been used to evaluate silver fir *de novo* assembled 454 transcriptome. A total of 3500 EST-SSRs were detected in the 454 transcriptome of silver fir. Most abundant are tri-nucleotide SSRs being followed by tetra- SSRs and di- SSRs. In addition, was determined the density, frequency, average length and average repeat number of EST-SSRs in the 454 transcriptome of silver fir.

Keywords: *Abies alba, bioinformatics tools, genomic resources, transcriptome, EST-SSRs*

INTRODUCTION

Forest genomic resources may be regarded as very important tools to unravel past evolutionary processes of forest tree species and to understand and forecast their future response to environmental changes (Neale & Kremer, 2011).

The two main objectives in forest genomic research are the breeding and genetic improvement of few tree species and sustainable management of forest populations by studying and deciphering genomic basis of complex traits involved in the adaptive process of forest trees. (Neale & Kremer, 2011).

Forest tree genomic research has primarily restricted to seven genera *(Pinus, Picea, Pseudotsuga, Populus, Eucalyptus, Quercus and Castanea)* and for some species of these genera, genome projects started with building reference genome sequence using next-generation sequencing (NGS) technologies (Neale & Kremer, 2011).

The first tree genome to be sequenced was for Black cottonwood *(Populus trichocarpa* Torr. & Gray*)* that has a relatively small genome size (450Mb) (Tuskan *et al.*, 2006)2006 compared to conifer tree species that are characterized by large genome size (e.g. *Picea abies* of 19.6 Gbp) (Neale *et al.*, 2013)2013, by low evolutionary rate of coding genes (Ritland, 2012) and by very large amount of repetitive DNA (Mackay *et al.*, 2012).

The first nuclear genome to be sequenced and assembled for conifer tree was for Norway spruce, one of the most widespread, economically and ecologically important tree species (Nystedt *et al.*, 2013). The Norway spruce genome assembly and expression data is publicly accessible through the ConGenIE database (http://congenie.org/).

The draft genomes of two other conifer species have been also published: white spruce (*Picea glauca*) (Birol *et al.*, 2013)2013 and loblolly pine (*Pinus taeda*) (Neale *et al.*, 2014)2014.

Assembling conifer species reference genome has become tractable only with the advent of next-generation sequencing (NGS) and advance

of assembly technologies. Ten conifer genome-sequencing projects are currently in progress in different sequencing research centres (Neale *et al.*, 2013)2013.

In addition to genome-sequencing projects in conifer forest tree species, there is also intensive progress in developing transcriptome resources and in resequencing for polymorphism discovery (Neale *et al.*, 2013)2013.

The whole genome sequencing and *de novo* assembly of large amount of short reads produced by NGS is much more challenging for non-model organisms, particularly for species characterized by large and repetitive genomes such as conifers when compared to transcriptome sequencing, being facilitated by the reduce amount of repetitive DNA and by increased coverage depth (Parchman *et al.*, 2010)2010.

The introduction of NGS has led to rapid increases in available transcriptome resources in conifer tree species, which are included in publicly accessible databases.

Available transcriptome resources are valuable for whole-genome assembly, marker discovery (e.g. EST-SSRs, SNPs), identification of structural genes, gene annotation, functional characterization of genes and for comparative phylogenetic analyses (Lorenz *et al.*, 2012)2012.

In this article, was described how different bioinformatics tools can be used for describing summary statistics of an assembled transcriptome and for identification and characterization of transcriptome derived EST-SSRs in silver fir (*Abies alba* Mill,).

MATERIALS AND METHODS

A normalized transcriptome of a 1-year-old *A. alba* seedling from the Black Forest (Forest District Calw, Germany, Voucher MB-P-001007, *Herbarium Marburgense,* University of Marburg) was sequenced on a 454 GsFLX Titanium platform (cDNA library preparation: Vertis Biotechnology AG, Freising, Germany, sequencing: Genoscreen, Lille, France). The 454 run yielded 1,521,698 reads with an average length of 359 nucleotide (nt) (Roschanski *et al.*, 2013)2013.

The Newbler Software Version 2.3 (454 Life Sciences) was used for trimming and *de novo* assembly of the raw reads that generated 22,561 contigs with the total assembly length of 12,131,455 base pairs. The minimum contig length was between 100 nucleotides and the maximum contig length was 2394 nucleotides with an average contig length of 537.72 nucleotides (Tab. 1).

Contigs were submitted to the transcriptome shotgun assembly database (TSA) at NCBI (Accession Numbers JV134525 to JV157085) (Roschanski *et al.*, 2013)2013.

The transcriptome statistics for total number of contigs, contig lengths, statistics for bases in the contigs, number of contigs in N50 and N90 were determined using the perl script, contig-stats.pl (available at http://milkweedgenome.org).

To evaluate the contig length distributions, the number of contigs in N25 and N75 and the total GC content, was used the perl script count_fasta.pl (available at http://wiki.bioinformatics.ucdavis.edu/index.php/Count_fasta.pl).

The software SSR Locator (Da Maia *et al.*, 2008)2008 was used to identify the transcriptome EST-SSRs with a minimum of 12 bp EST-SSR repeats: di-SSRs (x6), tri-SSRs (x4), tetra-SSRs (x3), penta-SSRs(x3) and hexa-SSRs (x3).

The density was calculated as number of base pairs per mega base pairs (bp/Mbp) of analyzed sequences, namely the length of detected SSRs out of the total length of the sequences for detection.

RESULTS AND DISCUSSIONS

The N50 score, which represents the length of the shortest contig at 50% of *Abies alba* transcriptome, is equal to 549 base pairs. The *Abies alba* transcriptome has a GC content of 42.19% in 5,117,968 nucleotides (Tab. 1, Fig. 1).

In total 3500 putative EST-SSRs were detected. The observed frequency of EST-SSRs in Abies alba is 0.45%, with an overall distribution density of 288.51 SSRs per Mb (Tab. 2).

Tri-nucleotide SSRs are most predominant in the *Abies alba* transcriptome being followed by tetra- SSRs and di- SSRs.

The analysis of the classified EST-SSRs by type and motif were clustered together (e.g., ACT=CTA=TCA=TGA=GAT=AGT), following (Jurka & Pethiyagoda, 1995), where repeat motifs were grouped according to circular permutations and/or reverse complements of each other (Tab. 3).

Among all dinucleotide repeats, the repeat type AG/GA/CT/TC (33.88 SSR/Mbp, 650.384 bp/Mbp) is the most frequent, while the repeat type GC/CG is absent.

Tab. 1. *Abies alba de novo* assembly transcriptome characteristics

Number of Contigs	22,561
Total assembly length (bp)	12,131,455
Minimum contig length (bp)	100
Maximum contig length (bp)	2,394
Average contig length (bp)	537.72
N25 contig length (nr. contigs used in calculus)	755 (3122 contigs)
N50 contig length (nr. contigs used in calculus)	549 (7876 contigs)
N50 (bp)	6,065,731
N75 contig length (nr. contigs used in calculus)	448 (14040 contigs)
N90 contig length (nr. contigs used in calculus)	377 (18429 contigs)
GC (%)	42.19%
Total GC count (bp)	5,117,968

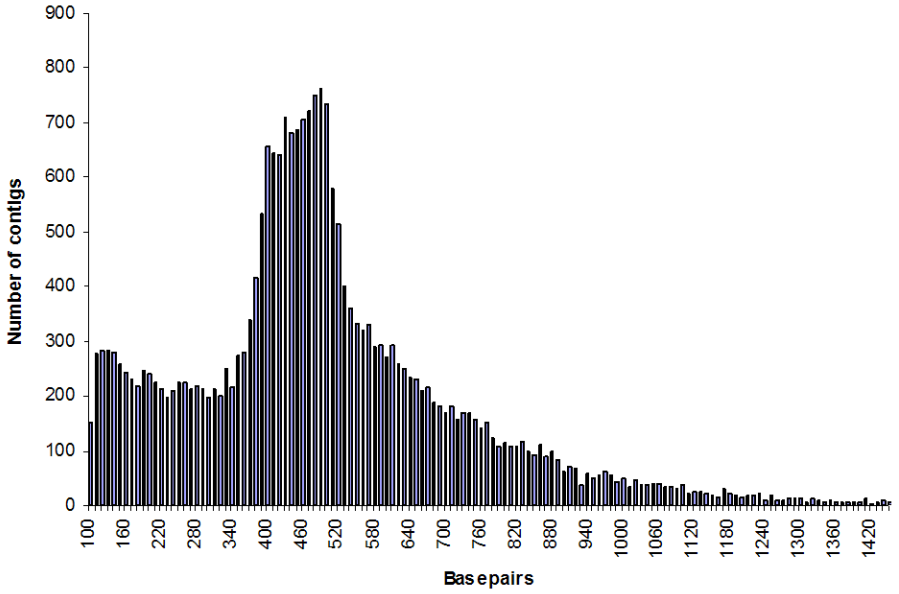

Fig. 1. Contig length distributions in silver fir *de novo* assembled 454 transcriptome

Among all trinucleotide repeats, the most represented ones were the repeat type ATG/TGA/GAT/CAT/ATC/TCA (36.76 SSR/Mbp, 576.19 bp/Mbp) (Tab. 3).

Transcriptome derived EST-SSRs markers are easily transferable to other species due to the higher level of sequence conservation of transcribed DNA across species (Varshney *et al.*, 2005; Zalapa *et al.*, 2012)2005; Zalapa<style face="italic"> *et al.*</style>, 2012.

The main drawbacks of transcriptome derived EST-SSRs markers are the expected lower genetic polymorphism and the concern that the selection pressure may bias the assessment of population genetic parameters (Ellis & Burke, 2007; Kim *et al.*, 2008)2008.

Recently, have been published (Postolache *et al.*, 2014)2014 16 new transcriptome derived EST-SSRs assembled in two 8-plexes based on the analysis of *Abies alba* transcriptome (Roschanski *et al.*, 2013)2013.

The cross-transferability of 16 EST-SSRs was 100% for eight Mediterranean firs, 81% for Asian firs and 75% for North American firs (Postolache *et al.*, 2014)2014.

Tab. 2. EST-SSRs in *Abies alba* transcriptome

EST-SSRs type	Number of EST-SSRs	Frequency %	Average repeat **number**	**Average Length (bp)**	**Density (SSRs/Mb)**
Dinucleotide	635	18.14	9.14	18.29	52.34
Trinucleotide	1484	42.40	4.79	14.36	122.33
Tetranucletide	844	24.11	3.36	13.42	69.57
Pentanucleotide	229	6.54	3.24	16.20	18.88
Hexanucleotide	308	8.80	3.38	20.28	25.39
Total	3500	100.00	4.78	16.51	288.51

Tab. 3. Frequency of EST-SSRs motifs in the *Abies alba* transcriptome

	Repeat motif	**Number of EST-SSRs**	**Frequency (%)**
	Dinucleotides	**635**	**18.14**
1	AT/TA	173	4.94
2	AG/GA/CT/TC	411	11.74
3	AC/CA/TG/GT	51	1.46
4	GC/CG	0	0.00
	Trinucleotides	**1484**	**42.40**
1	AAG/AGA/GAA/CTT/TTC/TCT	284	8.11
2	AAT/ATA/TAA/ATT/TTA/TAT	161	4.60
3	ATG/TGA/GAT/CAT/ATC/TCA	446	12.74
4	AAC/ACA/CAA/GTT/TTG/TGT	73	2.09
5	ACC/CCA/CAC/GGT/GTG/TGG	67	1.91
6	AGG/GGA/GAG/CCT/CTC/TCC	185	5.29
7	AGT/GTA/TAG/ACT/CTA/TAC	14	0.40
8	AGC/GCA/CAG/GCT/CTG/TGC	202	5.77
9	ACG/CGA/GAC/CGT/GTC/TCG	21	0.60
10	GGC/GCG/CGG/GCC/CCG/CGC	31	0.89

The newly developed transcriptome-derived EST-SSRs have been successfully used in new research studies, such as investigating effective gene flow (Leonarduzzi *et al.*, 2016)2016 and also in inferring population structure and demographic history in *Abies alba* natural populations from Italy and Eastern Europe (Romania and Bulgaria) (Postolache *et al.*, 2016; Piotti *et al.*, 2017).

Transcriptome resources have been also successfully used in transcriptome-wide association studies (TWAS) that are important in the context of global climate change as they can provide valuable information and support for forest management practices, by studying populations adaptation to climatic changes in space with the main goal to forecast populations response to climatic changes (Yeaman *et al.*, 2014)2014.

CONCLUSION

Developing genomic resources for non-model species based on high-throughput sequencing technologies is a feasible strategy more than ever before, in terms of cost-effectiveness for obtaining high-quality genome-scale data.

Newly developed genomic resources in silver fir can now be used in a wide range of studies, from estimating species demographic history and

population genetic structure to uncovering signals of local adaptation.

The newly developed genomic resources represent the starting point to depict the genomic basis of local adaptation in non-model forest species, such as conifer species that are characterized with mega-genome size and where developing transcriptome resources is considered most feasible strategy and cost-effective.

REFERENCES

1. Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Saint Yuen MM, Keeling CI, Brand D, Vandervalk BP (2013). Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics 29 (12): 1492-1497.
2. Da Maia LC, Palmieri DA, De Souza VQ, Kopp MM, de Carvalho FIF, Costa de Oliveira A (2008). SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. International journal of plant genomics. vol. 2008, Article ID 412696, 9 pages, 2008. doi:10.1155/2008/412696.
3. Ellis J, Burke J (2007). EST-SSRs as a resource for population genetic analyses. Heredity 99:125-132.
4. Jurka J, Pethiyagoda C (1995). Simple repetitive DNA sequences from primates: compilation and analysis. Journal of molecular evolution 40:120-126.
5. Kim KS, Ratcliffe ST, French BW, Liu L, Sappington TW (2008). Utility of EST-derived SSRs as population genetics markers in a beetle. Journal of Heredity 99:112-124.
6. Leonarduzzi C, Piotti A, Spanu I, Vendramin GG (2016). Effective gene flow in a historically fragmented area at the southern edge of silver fir (*Abies alba* Mill.) distribution. Tree Genetics & Genomes 12:95. doi:10.1007/s11295-016-1053-4.
7. Lorenz WW, Ayyampalayam S, Bordeaux JM, Howe GT, Jermstad KD, Neale DB, Rogers DL, Dean JF (2012). Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species. Tree Genetics & Genomes 8(6): 1477-1485.
8. Mackay J, Dean JF, Plomion C, Peterson DG, Canovas FM, Pavy N, Ingvarsson PK, Savolainen O, Guevara MA, Fluch S, Vinceti B, Abarca D, Diaz-Sala C, Cervera MT (2012). Towards decoding the conifer giga-genome. Plant molecular biology 80: 555-69.
9. Neale DB, Kremer A (2011). Forest tree genomics: growing resources and applications. Nature reviews. Genetics 12:111-122.
10. Neale DB, Langley CH, Salzberg SL, Wegrzyn JL (2013). Open access to tree genomes: the path to a better forest. Genome Biol 14:120.
11. Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, Martínez-García PJ, Vasquez-Gross HA, Lin BY, Zieve JJ, Dougherty WM, Fuentes-Soriano S, Wu L-S, Gilbert D, Marçais G, Roberts M, Holt C, Yandell M, Davis JM, Smith KE, Dean JF, Lorenz WW, Whetten RW, Sederoff R, Wheeler N, McGuire PE, Main D, Loopstra CA, Mockaitis K, de Jong PJ, Yorke JA, Salzberg SL, Langley CH (2014). Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome biology 15: R59.
12. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hallman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Kaller M, Luthman J, Lysholm F, Niittyla T, Olson A, Rilakovic N, Ritland C, Rossello JA, Sena J, Svensson T, Talavera-Lopez C, Theissen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalerao R, Bohlmann J, Bousquet J, Garcia Gil R, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, Van de Peer Y, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S (2013). The Norway spruce genome sequence and conifer genome evolution. Nature 497:579-584.
13. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC genomics 11:180.
14. Piotti A, Leonarduzzi C, Postolache D, Bagnoli F, Spanu I, Brousseau L, Urbinati C, Leonardi S, Vendramin GG (2017). Unexpected scenarios from Mediterranean refugial areas: disentangling complex demographic dynamics along the Apennine distribution of silver fir. Journal of Biogeography, doi.org/10.1111/jbi.13011.
15. Postolache D, Popescu F, Pitar D, Apostol E, Iordan A, Avram A, Iordan O, Zhelev P (2016). Origin, evolution and genetic structure of Silver fir stands of Romania evaluated through molecular markers. Revista de Silvicultură si Cinegetică 21:8-14.
16. Postolache D, Leonarduzzi C, Piotti A, Spanu I, Roig A, Fady B, Roschanski A, Liepelt S, Vendramin GG (2014). Transcriptome versus Genomic Microsatellite Markers: Highly Informative Multiplexes for Genotyping *Abies alba* Mill. and Congeneric Species. Plant Molecular Biology Reporter 32:750-760.
17. Ritland K (2012). Genomics of a phylum distant from flowering plants: conifers. Tree Genetics & Genomes 8: 573-582.
18. Roschanski AM, Fady B, Ziegenhagen B, Liepelt S (2013). Annotation and re-sequencing of genes from de novo transcriptome assembly of *Abies alba* (*Pinaceae*). Applications in plant sciences, 1(1).
19. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596-1604.
20. Varshney RK, Graner A, Sorrells ME (2005). Genic microsatellite markers in plants: features and applications. TRENDS in Biotechnology 23:48-55.
21. Yeaman S, Hodgins KA, Suren H, Nurkowski KA, Rieseberg LH, Holliday JA, Aitken SN (2014).

Conservation and divergence of gene expression plasticity following c. 140 million years of evolution in lodgepole pine (*Pinus contorta*) and interior spruce (*Picea glauca*× *Picea engelmannii*). New Phytologist 203:578-591.

22. Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2012). Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American journal of botany 99:193-208.