Tree Genetics & Genomes (2023) 19:14
https://doi.org/10.1007/s11295-023-01589-8

SHORT COMMUNICATION

Genome-wide SNP discovery in native American and Hungarian *Robinia pseudoacacia* genotypes using next-generation double-digest restriction-site-associated DNA sequencing (ddRAD-Seq)

Zoltán A. Köbölkuti[1,2] · Endre Gy. Tóth[1] · Zsolt Keserű[1] · John H. Fike[3] · Bence Bolla[1] · Tamás Ábri[1,4] · Attila Borovics[1] · Attila Benke[1] · Klára Cseke[1]

Received: 5 August 2022 / Revised: 22 January 2023 / Accepted: 29 January 2023 / Published online: 9 February 2023

Abstract

Robinia pseudoacacia L. (commonly known as black locust) is an economically and environmentally important plant, native to the eastern USA, and introduced into several European countries, including Hungary. An early successional leguminous tree, the black locust is characterized by tolerance to degraded sites, rapid growth rate, dense and rot-resistant wood, and heavy flowering. Due to its economic potential and environmental impact, the historic Hungarian breeding strategy targeted not only increased wood production but also in wood and honey-production quality. However, because many important features of the species are under polygenic control, genome-wide genetic data provided by high-throughput sequencing technology could make possible the identification of gene variants with identifiable functional effects on complex traits. Furthermore, the evaluation of the breeding efforts carried out so far would be also achievable, by comparing bred/selected genotypes with those from the original habitat. This paper provides a genomic dataset with highly variable SNP markers from native American and Hungarian *Robinia pseudoacacia* L. individuals. These SNP loci can be used to assess genetic differentiation, and to detect signatures of polygenic determination of economically important traits, providing a basis for further research into this species.

Keywords Single nucleotide polymorphism · Double digest RAD-Seq · Black locust · Genetic variation

Introduction

Native to the southeastern areas of North America (Boring and Swank 1984; Stone 2009; Li et al. 2014; Martin 2019), *R. pseudoacacia*'s biology, ecology, economic potential,

Communicated by: M. Troggio

✉ Zoltán A. Köbölkuti
kobolkuti.zoltan@uni-sopron.hu

Endre Gy. Tóth
toth.endre@uni-sopron.hu

Zsolt Keserű
keseru.zsolt@uni-sopron.hu

John H. Fike
jfike@vt.edu

Bence Bolla
bolla.bence@uni-sopron.hu

Tamás Ábri
abri.tamas@uni-sopron.hu

Attila Borovics
borovics.attila@uni-sopron.hu

Attila Benke
benke.attila@uni-sopron.hu

Klára Cseke
cseke.klara@uni-sopron.hu

1 University of Sopron, Forest Research Institute (UoS-FRI), Várkerület 30/A, Sárvár, Hungary

2 Bavarian Office for Forest Genetics (AWG), Dept. of Applied Forest Genetics Research, Forstamtsplatz 1, Teisendorf, Germany

3 School of Plant and Environmental Sciences, Virginia Tech, 185 Ag Quad Lane, 365 Smyth Hall, Blacksburg, VA, USA

4 University of Debrecen (UD), Faculty of Agricultural and Food Sciences and Environmental Management, Böszörményi U. 138, Debrecen, Hungary

and environmental impacts have been extensively studied in the USA, Europe, and China resulting in several reviews (Cierjacks et al. 2013; Sitzia et al. 2016; Vítková et al. 2017; Nicolescu et al. 2018, 2020; Puchałka et al. 2021). It is equally clear from these works, that black locust is a fast-growing tree species with the ability to fix nitrogen. Probably that is the reason why it is a popular plant used in vegetation restoration in degraded areas (Yüksek and Yüksek 2011), thereby becoming important for the development of a better ecological surrounding (Shi et al. 2021). Besides, it is highly appreciated in the wood industry producing large amounts of dense wood, highly resistant to decay being an excellent material for various sawn wood products (Keresztesi 1988; Nicolescu et al. 2020), also possessing a high combustion potential (Nicolescu et al. 2018), important as short rotation energy crop (Straker et al. 2015; Rédei et al. 2010) and finally, is used in biotherapy, apiculture, and food industries (Beldeanu 2008).

Introduced in Europe in the early seventeenth century (Boring and Swank 1984), its appearance in Hungary dates from 1710–1720 (Keresztesi 1983). In the country, the species spread rapidly due to its ability to utilize a wide range of sites, and in the 1960s, Hungary already possessed more black locust forests than all other European countries combined (Keresztesi 1980). It became one of the most suitable species for establishing environmentally valuable recreational plantations, and nowadays, black locust stands provide approximately 21% of the annual timber supply of Hungary (FAO 2020).

Considering all these previously mentioned ecological, economic, and environmental benefits but also noting that the black locust is amongst the 100 most invasive alien species on the European continent (Vilà et al. 2009), the improvement and management of stands are of paramount importance. Selection in Hungary targeted the production of high-quality trees, an increase in wood production quantity and in wood and honey-production quality (Redei et al. 2008; Nicolescu et al. 2018; Keserű et al. 2021; Ábri et al. 2022). However, breeding by crossing requires persistent work over several decades as the black locust is a heterozygotic species with many features with polygenic background. For that reason, the results of crossbreeding are influenced by chance. With no prior information about the genome, the development of the RADseq technology raised the hope that it would be possible to identify single polymorphic variants with identifiable functional effects on complex traits. However, such a claim requires extensive genetic investigation, since the enzyme selection, sequencing type (paired-end vs. single end), hence the presence of polymorphisms in restriction sites, the possible decrease of coverage, the sequencing errors (Verdu et al. 2016), biases of the de novo assembly, and the data missingness censoring strategy (Tripp et al. 2017) may affect the development of single nucleotide polymorphism markers.

With an aim to reveal patterns of genetic variation at the genome-wide scale, in this study, we performed double digest Restriction Site-Associated DNA sequencing (ddRAD-Seq) to generate unbiased reduced representation libraries of several *R. pseudoacacia* complete genomes. This dataset provides the opportunity to study not only the genetic background of polygenic traits but also can reveal important insights into the genetic differences between individuals from native American and Hungarian genotypes.

Material, methods, and results

Plant material

The laboratory phase of the experiment was conducted with a single sample/individual from 48 genotypes (Table S1). As one of the main aims was to compare individuals from native American populations with genotypes introduced in Hungary, the sampling was designed to include 16 North American genotypes, all from native habitats. The Hungarian sample series was divided into two major groups. One group with a total of 24 genotypes included old cultivars (Keresztesi 1988) and new candidates with excellent trunk quality, selected for breeding purposes. The second group contained eight samples collected either random, or deliberately from particularly old specimens, all with seedling origin. Thus, the "Bábolna black locust," registered as the oldest Hungarian tree from this species, an old specimen situated in front of the Hungarian Academy, and an individual of a similar age from Uzsa (Hungary) is also present in the study.

Library preparation and RAD-tag sequencing

Total genomic DNA extraction was performed from leaves per each individual using the ATMAB method (Dumolin et al. 1995) with minor modifications. The Qubit dsDNA BR Assay Kit and Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) were used to quantify the extracted DNA of each sample. A total of 50 ng of DNA per sample was double digested with *PstI* and *MspI* (FastDigest restriction enzymes; Thermo Fisher Scientific, Waltham, MA, USA). The enzyme combination was selected based on a preliminary restriction site analysis carried out in CLC Genomic Workbench version 12.0 (QIAGEN Bioinformatics, Hilden, Germany). Fragments were double-sided size selected using KAPA PureBeads (Roche, Basel, Switzerland), to isolate fragments in the range of 300–600 bp. Inserts were quantified (3 ng), then ligated to adapters (Table S2) by using T4 DNA Ligase according

to the manufacturers' protocol (Thermo Fisher Scientific, Waltham, MA, USA).

Purification of the ligated products was performed using 0.8 vol KAPA PureBeads (Roche, Basel, Switzerland) PCR amplification with NEBNext Multiplex Oligos for Illumina (Dual Index Set 1; New England Biolabs, Ipswich, MA, USA) and KAPA HiFi Hotstart Ready Mix (Roche, Basel, Switzerland). The amount of 0.5–0.5 µl of i5 and i7 indexed primers were used per reaction. Thermal cycling conditions were as follows: a 3-min initial denaturation at 95 °C; 17 cycles of 30 s of denaturation at 95 °C, 30 s of annealing at 55 °C, and 30 s extension at 72 °C; and a final 5-min extension at 72 °C. High Sensitivity DNA1000 ScreenTape system with 2200 Tapestation (Agilent Technologies, Santa Clara, CA, USA) and dsDNA HS Assay Kit with Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) was used to determine the quality and quantity of the amplicon library. Pooled libraries were diluted to 10 pM for 2×301 bp paired-end sequencing with 600-cycle sequencing kit v3.1 (Illumina, San Diego, CA, USA). Nucleotide sequences of the libraries were determined on a MiSeq Sequencing System (Illumina, San Diego, CA, USA) according to the manufacturer's protocol.

De novo assembly and SNP calling

All bioinformatic steps were performed on a Silicon Computers (SGI) HPC server, allocating 40 cores (80 threads) and 38 GB RAM, located at the University of Sopron, Sopron, Hungary.

Eight fastq files containing raw short read sequences (from four different runs: four R1 and four R2) of 48 samples each were joined resulting in cca. 96,00,000 reads. MiSeq Control Software (Illumina, San Diego, CA, USA) was used for demultiplexing and adapter-trimming of all sequences. Trimming the bases at the 3′ and the 5′ end with a quality score of less than 30 was performed with the implemented FastQ Toolkit. Reads having mean quality score less than 30 and shorter than 200 bp were filtered. As a next step, computational processing of short-read data was carried out with Stacks 2.0 (Catchen et al. 2013; Rochette et al. 2019). Reads were quality filtered for the second time using a sliding-window method (15% of read length) implemented in "process_radtags." Reads having a quality score below 90% (raw phred score of 10) were discarded (Catchen et al. 2011). RAD loci were reconstructed following the De novo pipeline implemented in "denovo_map.pl" command, in which "ustacks" build loci and call SNPs in each sample. Further, a catalogue of all loci for all the samples was created by "cstacks," then the match of loci of the samples against the catalogue was made by "sstacks" (Catchen et al. 2011; Rochette et al. 2019). Then, we optimized the parameters of the "denovo_map.pl" command by defining (*m*) the minimum number of reads to consider an allele, (*M*) the maximum number of mismatches allowed between two alleles, and (*n*) the maximum number of mismatches allowed between two individual loci to consider them as homologous (Mastretta-Yanes et al. 2015; Paris et al. 2017). These parameters were optimized throughout the "r80" method, by effectively maximizing the number of polymorphic loci found in 80% of the individuals (Paris et al. 2017). The iterative values (ranging from 2 to 10) of *M* and *n* were investigated by how they affect the gained polymorphic loci, then the thresholds of $M = 3$, $n = 3$ were chosen, applying the $M = n$ rule for the final run (Paris et al. 2017). The default $m = 3$ value (three identical reads) was selected for the stack-dept parameter. Assembly of paired-end contigs and re-calling the SNPs using the population-wide data (Rochette et al. 2019) was used by "gstacks." Then by "denovo_map.pl" pipeline, we aligned paired-end reads for a total of 210,491 RAD loci, composed of 31,650,195 sites. One hundred forty-seven of these sites were filtered, 220,105 variant sites remained. The mean genotyped sites per locus were 150.36 bp (stderr 0.00). The depth of sequencing coverage is presented in Table S3 and Fig S1; the number of reads incorporated for each sample processed as calculated by "ustacks," in Table S3.

With the "populations" program, we called the set of SNP genotypes. Further processing was accomplished with the VCFtools program package (Danecek et al. 2011). After filtering genotypes called below 50% across all individuals and SNPs that have a minor allele count less than 3, were kept all 48 individuals and 53,711 out of a possible of 220,105 SNP sites. Applying the next filter for a minimum depth for a genotype call and a minimum mean depth, all 53,711 sites were kept. A next step was performed to assess individual levels of missing data by quantifying missingness for each sample. To observe the degree of missingness on a per sample basis, the output file (Table S4) and the histogram (Figure S2) were examined, and based on the previous, a list of six individuals with more than 50% missing data was created, with the aim to identify these individuals in the dataset. By the depth of sequencing coverage (Table S3), these samples were in the upper coverage range of all samples, but still outside our acceptance threshold as missing.

Following the removal of these individuals (ROB-HU-CULT13, ROB-HU-CULT15, ROB-HU-CULT4, ROB-HU-CULT5, ROB-HU-CULT6, ROB-HU-CULT7), we provided a dataset restricted to variants called in a high percentage of individuals. In a next step, we filtered by mean depth of genotypes. By applying a genotype call rate (95%) across all 42 individuals, our final dataset consisted of 112 SNP sites.

We validated our data with manual, visual, and numerical tests. The quality search of raw short read data of all individuals (48) was assessed using FastQC v0.11.9 (Andrews 2010) by investigating per-base quality and per-sequence quality at three steps: at raw read state directly

after sequencing, after sequence quality trimming (3′ and 5′ end), and following sequence processing (whole read filtering). Results were consistent with a standard Illumina run, and therefore considered to be of sufficiently high quality for further analysis (Kircher et al. 2011).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11295-023-01589-8.

Acknowledgements The authors are grateful to Tamás Fonyó (UOS-FRI, Sopron, Hungary) for his assistance in providing computational resources. We also thank Zoltán Bihari (Xenovea Ltd.) for providing sequencing resources and useful comments on data analysis.

Author contribution KC, ABE, and ZAK conceived the study. ZK, JHF, BB, TÁ, and ABE contributed in sampling of plant materials. KC performed laboratory analysis. ZAK and EGT processed genomic data and performed statistical analyses and simulations. ZAK wrote the manuscript. All co-authors provided feedback on the manuscript drafts.

Funding Open access funding provided by University of Sopron. This article was made in frame of the project TKP2021-NKTA-43 which has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary (successor: Ministry of Culture and Innovation of Hungary) from the National Research, Development and Innovation Fund, financed under the TKP2021-NKTA funding scheme.

Declarations

Conflict of interest The authors declare no competing interests.

Data Archiving Statement The joined, demultiplexed and adapter-trimmed raw short read sequences (R1, R2) of 48 samples have been deposited in the NCBI Sequence Read Archive (SRA); BioProject ID: PRJNA924530 (http://www.ncbi.nlm.nih.gov/bioproject/924530). The full dataset including the 53,711 polymorphic sites (Rpseudo_ddRAD_unfiltered.vcf) and the filtered dataset (Rpseudo_ddRAD_filtered.vcf) is available from the ZENODO repository https://doi.org/10.5281/zenodo.7525264.

References

Ábri T, Keserű Z, Borovics A, Rédei K, Csajbók J (2022) Comparison of juvenile, drought tolerant black locust (Robinia pseudoacacia L.) clones with regard to plant physiology and growth characteristics in Eastern Hungary: early evaluation. Forests 13:292. https://doi.org/10.3390/f13020292

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Beldeanu EC (2008) Produse forestiere. Editura Universității Transilvania din Brașov, Brașov, p 331. (in Romanian)

Boring LR, Swank WT (1984) The role of black locust (*Robinia pseudo-acacia*) in forest succession. J Ecol 749–766. https://doi.org/10.2307/2259529

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3: Genes Genom Genet 1:171–182. https://doi.org/10.1534/g3.111.000240

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140. https://doi.org/10.1111/mec.12354

Cierjacks A, Kowarik I, Joshi J, Hempel S, Ristow M, von der Lippe M, Weber E (2013) Biological flora of the British Isles: *Robinia pseudoacacia*. J Ecol 10:1623–1640. https://doi.org/10.1111/1365-2745.12162

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dumolin S, Demesure B, Petit RJ (1995) Inheritance of chloroplast and mitochondrial genomes in pedunculate oak investigated with an efficient PCR method. Theor Appl Genet 91:1253–1256. https://doi.org/10.1007/bf00220937

Keresztesi B (1980) The black locust. Unasylva 32:23–33

Keresztesi B (1983) Breeding and cultivation of black locust, *Robinia pseudoacacia*, in Hungary. For Ecol Manag 6:217–244. https://doi.org/10.1016/s0378-1127(83)80004-8

Keresztesi B (1988) The black locust. Akadémiai Kiadó. Budapest. p 196. (in Hungarian)

Keserű Z, Borovics A, Ábri T, Rédei KM, Lee IH, Lim H (2021) Growing of black locust (*Robinia pseudoacacia* L.) candidate cultivars on arid sandy site. Acta Silv Et Lignaria Hung 17:51–61. https://doi.org/10.37045/aslh-2021-0004

Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of illumina sequencing data. BMC Genomics 12:382. https://doi.org/10.1186/1471-2164-12-382

Li G, Xu G, Guo K, Du S (2014) Mapping the global potential geographical distribution of black locust (*Robinia pseudoacacia* L.) using herbarium data and a maximum entropy model. Forests 5:2773–2792. https://doi.org/10.3390/f5112773

Martin GD (2019) Addressing geographical bias: a review of *Robinia pseudoacacia* (black locust) in the Southern Hemisphere. S Afr J Bot 125:481–492. https://doi.org/10.1016/j.sajb.2019.08.014

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC (2015) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol Ecol Resour 15:28–41. https://doi.org/10.1111/1755-0998.12291

Nicolescu VN, Hernea C, Bakti B, Keserű Z, Antal B, Rédei K (2018) Black locust (*Robinia pseudoacacia* L.) as a multipurpose tree species in Hungary and Romania: a review. J for Res 29:1449–1463. https://doi.org/10.1007/s11676-018-0626-5

Nicolescu VN, Rédei K, Mason WL, Vor T, Pöetzelsberger E, Bastien JC, Pástor M (2020) Ecology, growth and management of black locust (*Robinia pseudoacacia* L.), a non-native species integrated into European forests. J for Res 31:1081–1101. https://doi.org/10.1007/s11676-020-01116-8

Paris JR, Stevens JR, Catchen JM (2017) Lost in parameter space: a road map for stacks. Methods Ecol Evol 8:1360–1373. https://doi.org/10.1111/2041-210x.12775

Puchałka R, Dyderski MK, Vítková M, Sádlo J, Klisz M, Netsvetov M, Jagodziński AM (2021) Black locust (*Robinia pseudoacacia* L.)

range contraction and expansion in Europe under changing climate. Glob Change Biol 27:1587–1600. https://doi.org/10.1111/gcb.15486

Redei K, Osvath-Bujtas Z, Veperdi I (2008) Black locust (*Robinia pseudoacacia* L.) improvement in Hungary: a review. Acta Silv Lignaria Hung 4:127–132

Rédei K, Veperdi I, Tomé M, Soares P (2010) Black locust (*Robinia pseudoacacia* L.) short-rotation energy crops in Hungary: a review. Silva Lusit 18(2):217–223

Rochette NC, Rivera-Colón AG, Catchen JM (2019) Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. Mol Ecol 28:4737–4754. https://doi.org/10.1111/mec.15253

Shi Z, Bai Z, Guo D, Chen M (2021) Develop a soil quality index to study the results of black locust on soil quality below different allocation patterns. Land 10:785. https://doi.org/10.3390/land10080785

Sitzia T, Cierjacks A, de Rigo D, Caudullo G (2016) *Robinia pseudoacacia* in Europe: distribution, habitat, usage and threats. In: San-Miguel-Ayanz J, de Rigo D, Caudullo G, Houston Durrant T, Mauri A (eds) European atlas of forest tree species. Publication office of the European Union, Luxembourg, pp e014e79

Stone KR (2009) *Robinia pseudoacacia*. Fire Effects Information System. US Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory. http://www.fs.fed.us/database/feis/dostępon-line, 3, 2013

Straker KC, Quinn LD, Voigt TB, Lee DK, Kling GJ (2015) Black locust as a bioenergy feedstock: a review. Bioenergy Res 8:1117–1135. https://doi.org/10.1007/s12155-015-9597-y

Tripp EA, Tsai YE, Zhuang Y, Dexter KG (2017) RADseq dataset with 90% missing data fully resolves recent radiation of Petalidium (Acanthaceae) in the ultra-arid deserts of Namibia. Ecol Evol 7:7920–7936. https://doi.org/10.1002/ece3.3274

UN FAO. Global Forest Resources Assessment 2020 Report Hungary; UN FAO: Rome, Italy, 2020. https://www.fao.org/forest-resources-assessment/fra-2020/country-reports/en/

Verdu CF, Guichoux E, Quevauvillers S, De Their O, Laizet Y, Delcamp A, Gévaudant F, Monty A, Porté AJ, Lejeune P, Lassois L (2016) Dealing with paralogy in RAD seq data: in silico detection and single nucleotide polymorphism validation in *Robinia pseudoacacia* L. Ecol Evol 6:7323–7333. https://doi.org/10.1002/ece3.2466

Vilà M, Başnou C, Gollasch S, Josefsson M, Pergl J, Scalerra R (2009) One hundred of the most invasive alien species in Europe. In: DAISIE (ed) DAISIE handbook of alien species in Europe invading nature. Springer series in invasion ecology, vol 3. Springer, Dordrecht, pp 265–268. https://doi.org/10.1007/978-1-4020-8280-1_12

Vítková M, Müllerová J, Sádlo J, Pergl J, Pyšek P (2017) Black locust (*Robinia pseudoacacia*) beloved and despised: a story of an invasive tree in Central Europe. For Ecol Manag 384:287–302. https://doi.org/10.1016/j.foreco.2016.10.057

Yüksek T, Yüksek F (2011) The effects of restoration on soil properties in degraded land in the semi-arid region of Turkey. CATENA 84:47–53. https://doi.org/10.1016/j.catena.2010.09.002