HAL
open science

RNAseq based variant dataset in a black poplar association panel

Odile Rogier, Aurélien Chateigner, Marie-Claude Lesage-Descauses, Claire Mandin, Véronique Brunaud, José Caius, Ludivine Soubigou-Taconnat, José Almeida-Falcon, Catherine Bastien, Vanina Benoit, et al.

HAL Id: hal-04225417
https://hal.inrae.fr/hal-04225417

Submitted on 2 Oct 2023

DATA NOTE

RNAseq based variant dataset in a black poplar association panel

Odile Rogier[1], Aurélien Chateigner[1], Marie-Claude Lesage-Descauses[1], Claire Mandin[1], Véronique Brunaud[2,3], José Caius[2,3], Ludivine Soubigou-Taconnat[2,3,4], José Almeida-Falcon[5], Catherine Bastien[1], Vanina Benoit[1], Guillaume Bodineau[6], Nathalie Boizot[1], Corinne Buret[1], Jean-Paul Charpentier[1], Annabelle Déjardin[1], Alain Delaunay[7], Régis Fichot[7], Véronique Laine Prade[1], Françoise Laurans[1], Isabelle Le Jan[7], Anne-Laure Legac[7], Stéphane Maury[7], Mesfin Nigussie Gebreselassie[1], Patrick Poursat[6], Céline Ridel[1], Léopoldo Sanchez[1], Véronique Jorge[1] and Vincent Segura[1,8]*

Abstract

Objective Black poplar (*Populus nigra* L.) is a species native to Eurasia with a wide distribution area. It is an ecologically important species from riparian ecosystems, that is used as a parent of interspecific (*P. deltoides* x *P. nigra*) cultivated poplar hybrids. Variant detection from transcriptomics sequences of 241 *P. nigra* individuals, sampled in natural populations from 11 river catchments (in four European countries) is described here. These data provide new valuable resources for population structure analysis, population genomics and genome-wide association studies.

Data description We generated transcriptomics data from a mixture of young differentiating xylem and cambium tissues of 480 *Populus nigra* trees sampled in a common garden experiment located at Orléans (France), corresponding to 241 genotypes (2 clonal replicates per genotype, at maximum) by using RNAseq technology. We launched on the resulting sequences an in-silico pipeline that allowed us to obtain 878,957 biallelic polymorphisms without missing data. More than 99% of these positions are annotated and 98.8% are located on the 19 chromosomes of the *P. trichocarpa* reference genome. The raw RNAseq sequences are available at the NCBI Sequence Read Archive SPR188754 and the variant dataset at the *Recherche Data Gouv* repository under https://doi.org/10.15454/8DQXK5.

Keywords *Populus nigra*, RNA polymorphisms, SNV discovery, Genotyping, Bioinformatics pipeline, Transcriptomics

---

*Correspondence:
Vincent Segura
vincent.segura@inrae.fr

[1] INRAE, ONF, BioForA, 45075 Orléans, France
[2] Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91190 Gif sur Yvette, France
[3] Université Paris Cité, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91190 Gif sur Yvette, France
[4] Groupe d'Étude et de Contrôle des Variétés et des Semences (GEVES), Station Nationale d'Essais de Semences (SNES), 49071 Beaucouzé, France
[5] INRAE, InfoSol, 45075 Orléans, France
[6] INRAE, GBFor, 45075 Orléans, France
[7] Laboratoire de Biologie des Ligneux et des Grandes Cultures, INRAE, EA1207 USC1328, Université d'Orléans, 45067 Orléans, France
[8] UMR AGAP Institut, CIRAD, INRAE, Institut Agro Montpellier, Univ Montpellier, 34398 Montpellier, France

Rogier *et al. BMC Research Notes*   (2023) 16:248

Page 2 of 5

Objective

Of the twenty-nine species in the genus *Populus*, black poplar (*P. nigra* L.) is native to Eurasia with a wide distribution area including Europe, as well as the southwest and central Asia, and northwest Africa [1]. It is regarded as a keystone species for riparian ecosystems in ecological and conservation studies [2] and it has an interest as a parental pool in interspecific (*P. deltoides* x *P. nigra*) poplar breeding programs as the origin of cultivated hybrids [3].

In this study, we selected 241 genotypes from a *P. nigra* collection of 587 genotypes previously genotyped with an Illumina 12 K Infinium Bead-Chip array (8000 Single Nucleotide Variants (SNVs); [4]), which in turn belong to a larger collection of 1098 cloned genotypes sampled in natural populations from 11 river catchments [5] in four European countries. These 241 genotypes were previously studied for wood properties in 2 sites (Savigliano-2011 & Orléans-2012; [6]), and their selection was based on the following set of criteria defined following a first analysis of population structure with 8000 SNVs: (i) introgression < 10% of the worldwide-spread fastigiated form *P. nigra* var *italica*, (ii) proportion of recruitment to their ancestral population > 50% and (iii) survival in the common garden located at Orléans, France.

We generated transcriptomics data from young differentiating xylem and cambium tissues of these *P. nigra* selected genotypes by using RNAseq technology. We launched on the resulting sequences an in-silico pipeline [7] that allowed us to obtain 878,957 polymorphisms. We already used this data in Chateigner et al. [8] and in Wade et al. [9] for phenotype prediction. These data provide new valuable resources for a wide variety of genome-based studies, ranging from population structure analysis over distribution ranges, to genomic prediction and genome-wide association studies (GWAS) for traits related to wood properties and growth, for example.

Data description

Young differentiating xylem and cambium tissues were harvested in June 2015 from 480 *Populus nigra* trees from a common garden located at Orléans (France) (241 genotypes, 2 clonal trees per genotype, Data file 1 [10], Data set 1 [11]). RNA from the xylem and cambium were extracted with RNeasy Plant kit (Qiagen, France) according to manufacturer's recommendations. Treatment with DNase I (Qiagen, France) was carried out to ensure elimination of genomic DNA. RNA was eluted in RNAse-DNAse free water and quantified with a Nanodrop spectrophotometer. RNA from xylem and cambium of the same plant were pooled in an equimolar extract (250 ng/µL) and sent to the sequencing platform. The sequencing platform POPS (transcriptOmic Platform of Institute of Plant Sciences—Paris-Saclay) prepared the RNAseq libraries from polyA-RNA selection using the TruSeq_Stranded_mRNA_SamplePrep_Guide_15031047_D protocol (Illumina, California, U.S.A.).

Identity of each sample was checked with the following procedure: a first round of variant detection was performed with FreeBayes (v.1.0.0) [12], then IBS (Identity By State) was calculated between samples of the same individual, as well as between the samples of the present study and the 852 individuals that had been previously genotyped with the Illumina 12 K Infinium Bead-Chip array [4] (Data file 2 [13]). After removing sampling errors and correcting identities, 241 unique genotypes were considered corresponding to 461 FASTQ files (Data set 2 [14]).

The sequences for each sample were processed with the pipeline defined in Rogier et al. [7] with small modifications to detect SNVs. All experiment steps (from growth conditions to bioinformatic analyses) are available in the CATdb database (Data set 3 [15]). Briefly, the reads were first trimmed with Trimmomatic (v.0.38) [16] to remove adapter and low-quality sequences. Then, they were aligned to the *Populus trichocarpa* reference genome v.3.0 [17] using the BWA-MEM algorithm (v.0.7.12) [18]. We followed the GATK Best Practices [19, 20] for RNAseq short variant discovery: we first marked the duplicates with the MarkDuplicates from the Picard tools (v.2.0.1) [21] and then used the SplitNCigarReads, the Indel Realignment and the Base Quality Recalibration tools from GATK (v.3.5) [22]. SNV and short insertions and deletions were genotyped for all the sequenced trees (the same genotypes were pooled together) with 3 variants callers: (i) GATK using the HaplotypeCaller tool in single-sample calling mode followed by joint genotyping of the samples with the GenotypeGVCFs tool; (ii) Free-Bayes (v.1.0.0) [12] in a multi-sample mode and (iii) the mpileup tool from SAMtools (v.1.3.1) [23] in a multi-sample mode followed by bcftools (v.1.3.1) [24].

The resulting 3 files (one per caller) were filtered with VCFtools (v.0.1.15) [25] to obtain only the intra-specific (*Populus nigra*) biallelic SNV with a variant quality score (QUAL) threshold over 30. These filtered files were then combined together with the vcf-isec tool from VCFtools and we only kept the genotype calls that were detected by at least 2 variant callers. Otherwise, the genotype call was set as a missing value for this particular individual. As there remained some missing data (Data file 3 [26]), genotype imputation was performed using the Fimpute program (v.2.2) [27] for the SNV located on the chromosomes/scaffolds that contain at least 2 SNV. Thereby, 9.73% of missing values were imputed. This yielded genotypes at 878,957 biallelic sites for the 241 *Populus nigra* individuals (Data file 4 [28]) without missing data.

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	collection_POPULUS_NIGRA_RNASEQ_PANEL *(list of the studied genotypes)*	Tabular text file (.tab)	*Recherche Data Gouv* repository: https://doi.org/10.57745/GKXDSQ; [10]
Data set 1	Information on the studied genotypes	Web files (.html)	GnpIS: collection:POPULUS_NIGRA_RNASEQ_PANEL; [11]
Data file 2	SNV_POPULUS_NIGRA_RNASEQ_PANEL_quality_control *(Quality control figures)*	PDF file (.pdf)	*Recherche Data Gouv* repository: https://doi.org/10.57745/SSDFV2; [13]
Data set 2	Raw RNA-seq files	Sequencing files (.fastq)	NCBI Sequence Read Archive: SPR188754; [14]
Data set 3	Steps of the experiment, from growth conditions to bioinformatic analyses	Web files (.html)	CATdb: experiment = 640; [15]
Data file 3	SNV_raw_POPULUS_NIGRA_RNASEQ_PANEL *(Initial variant calling file)*	Variant Call Format file (.vcf)	*Recherche Data Gouv* repository: https://doi.org/10.57745/RBR6X0; [26]
Data file 4	SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL *(Imputated variant calling file)*	Variant Call Format file (.vcf)	*Recherche Data Gouv* repository: https://doi.org/10.57745/5IQLI9; [28]
Data file 5	SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL.variant_function *(Annotation file1)*	Tabular text files (.tab)	*Recherche Data Gouv* repository: https://doi.org/10.57745/PAEKL7; [29]
Data file 6	SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL.exonic_variant_function *(Annotation file2)*	Tabular text files (.tab)	*Recherche Data Gouv* repository: https://doi.org/10.57745/EG9HOE; [30]
Data file 7	SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL_analysis_figures *(Analysis figures)*	PDF file (.pdf)	*Recherche Data Gouv* repository: https://doi.org/10.57745/BTBR; [32]

878,893 of them have an annotation (Data file 5 [29], Data file 6 [30]) from ANNOVAR (v.2017Jul16) [31] (Data file 7 [32]). Among them, 868,861 are located on the 19 chromosomes (Data file 7, [32]). A total of 26,909 genes harbored between 1 and 309 SNVs.

Limitations

- The SNVs are limited to expressed genes: 26,909 vs a maximum of 41,335 *P. trichocarpa* protein-coding genes. We found a correlation between expression level and SNV density at gene level, which is likely due to an increase in coverage for highly expressed genes (Data file 7 [32]).
- RNA were extracted from young differentiating xylem and cambium tissues, which might not be representative of transcriptional activity of other tissues of the tree. The found SNVs are therefore specific to the genes expressed in these two tissues. Nevertheless, this makes this SNV dataset clearly appropriate to study the genetics of wood formation.
- The SNVs are limited to the gene space: 92.55% of the found SNV are exonic, intronic, in the 3'UTR (UnTranslated Region) or in the 5'UTR (Data file 7 [32]).

Abbreviations

GATK	Genome analysis ToolKit
GWAS	Genome-wide association studies
IBS	Identity by state
mRNA	Messenger RiboNucleic Acid
NCBI	National center for biotechnology information
RNA	RiboNucleic acid
RNAseq	RiboNucleic acid sequencing
SNV	Single nucleotide variant
UTR	UnTranslated region

Availability of data and materials
The raw RNAseq sequences are available in the National Center for Biotechnology Information (NCBI) Sequence Read Archive with the SPR188754 identifier [14]. The 241 associated accessions can be found in the GnpIS Information System under the collection "POPULUS_NIGRA_RNASEQ_PANEL" [11] or with the DOIs in the table of the studied genotypes [10]. The steps of the experiment (from growth conditions to bioinformatic analyses) are described in the CATdb database [15]. The other data described in this Data note can be freely and openly accessed on the *Recherche Data Gouv* repository under https://doi.org/10.15454/8DQXK5 [33]. Please see Table 1 and references [13, 26, 28–30, 32] for details and links to the data.

Rogier *et al. BMC Research Notes*   (2023) 16:248

Page 4 of 5

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

References

1. Dickmann DI, Kuzovkina J. Poplars and willows of the world, with emphasis on silviculturally important species. In: Isebrands JG, Richardson J, editors. Poplars and willows: trees for society and the environment. Wallingford: CABI; 2014.
2. Imbert E, Lefèvre F. Dispersal and gene flow of *Populus nigra* (*Salicaceae*) along a dynamic river system. J Ecol. 2003;91(3):447–56. https://doi.org/10.1046/j.1365-2745.2003.00772.x.
3. Stanton BJ, Serapiglia MJ, Smart LB. The domestication and conservation of *Populus* and *Salix* genetic resources. In: Isebrands JG, Richardson J, editors. Poplars and willows: trees for society and the environment. Wallingford: CABI; 2014.
4. Faivre-Rampant P, Zaina G, Jorge V, Giacomello S, Segura V, Scalabrin S, et al. New resources for genetic studies in *Populus nigra*: genome-wide SNP discovery and development of a 12k infinium array. Mol Ecol Resour. 2016;16(4):1023–36. https://doi.org/10.1111/1755-0998.12513.
5. Guet J, Fabbrini F, Fichot R, Sabatti M, Bastien C, Brignolas F. Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (*Populus nigra* L.). Tree Physiol. 2015;35(8):850–63. https://doi.org/10.1093/treephys/tpv056.
6. Gebreselassie MN, Ader K, Boizot N, Millier F, Charpentier JP, Alves A, et al. Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from *Populus nigra* (L.) natural populations. Ind Crops Prod. 2017. https://doi.org/10.1016/j.indcrop.2017.05.013.
7. Rogier O, Chateigner A, Amanzougarene S, Lesage-Descauses MC, Balzergue S, Brunaud V, et al. Accuracy of RNAseq based SNP discovery and genotyping in *Populus nigra*. BMC Genomics. 2018;19(1):909. https://doi.org/10.1186/s12864-018-5239-z.
8. Chateigner A, Lesage-Descauses MC, Rogier O, Jorge V, Leplé JC, Brunaud V, et al. Gene expression predictions and networks in natural populations supports the omnigenic theory. BMC Genomics. 2020;21(1):416. https://doi.org/10.1186/s12864-020-06809-2.
9. Wade AR, Duruflé H, Sanchez L, Segura V. eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction. BMC Genomics. 2022;23(1):476. https://doi.org/10.1186/s12864-022-08690-7.
10. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. "Collection_POPULUS_NIGRA_RNASEQ_PANEL.tab", Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project), Recherche Data Gouv, V2; 2022; https://doi.org/10.5774/GKXDSQ.
11. GnpIS: Genetic dans Genomic Information System. GnpIS; Collection: POPULUS_NIGRA_RNASEQ_PANEL. https://urgi.versailles.inrae.fr/faidare/search?gl=POPULUS_NIGRA_RNASEQ_PANEL
12. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv. 2012. https://doi.org/10.4855/arXiv.1207.3907.
13. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. "SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL_quality_control.pdf", Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project), Recherche Data Gouv, V2; 2022; https://doi.org/10.5774/SSDFV2

14. NCBI Sequence Read Archive. 2020; https://identifiers.org/ncbi/insdc.sra:SRP188754.
15. CATdb: a Plant Transcriptome Database; Available from: http://tools.ips2.u-psud.fr/CATdb/ficheexperiment.html?experiment=640;
16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
17. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science. 2006;313(5793):1596–604. https://doi.org/10.1126/science.1128691.
18. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. https://doi.org/10.48550/arXiv.1303.3997.
19. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8. https://doi.org/10.1038/ng.806.
20. Van der Auwera GA, O'Connor BD. Genomics in the cloud: using docker, GATK, and WDL in terra. 1st ed. Sebastopol: O'Reilly Media, Inc.; 2020.
21. Broad Institute. Picard Tools. Broad Institute, GitHub repository; 2018; https://broadinstitute.github.io/picard/.
22. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. https://doi.org/10.1101/gr.107524.110.
23. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10(2):008. https://doi.org/10.1093/gigascience/giab008.
24. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93. https://doi.org/10.1093/bioinformatics/btr509.
25. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330.
26. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. "SNV_raw_POPULUS_NIGRA_RNASEQ_PANEL.vcf", Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project), Recherche Data Gouv, V2; 2022; https://doi.org/10.5774/RBR6X0.
27. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15(1):478. https://doi.org/10.1186/1471-2164-15-478.
28. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. "SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL.vcf", Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project), Recherche Data Gouv, V2; 2022; https://doi.org/10.5774/5IQLI9.
29. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. "SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL.variant_function", Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project), Recherche Data Gouv, V2; 2022; https://doi.org/10.5774/PAEKL7.
30. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. "SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL.exonic_variant_function", Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project), Recherche Data Gouv, 2022; https://doi.org/10.5774/EG9HOE.
31. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164. https://doi.org/10.1093/nar/gkq603.
32. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. "SNV_imputated_POPULUS_NIGRA_RNASEQ_PANEL_analysis_figures.pdf", Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project). Recherche Data Gouv, 2022; https://doi.org/10.5774/BQQTBR.

Rogier *et al. BMC Research Notes*   (2023) 16:248

Page 5 of 5

33. Rogier O, Chateigner A, Wade AR, Lesage-Descauses MC, Brunaud V, Caius J, Soubigou-Taconnat L, Duruflé H, Sanchez L, Jorge V, Segura V. Phenotypic, genotypic and transcriptomic data of 241 Populus nigra (from the Sybiopop project). Recherche Data Gouv. 2022. https://doi.org/10.1545/8DQXK5.

Publisher's Note