The mixed liver and kidney transcriptome dataset of *Darevskia valentini* rock lizard

Sergei S. Ryakhovsky1,2*, Daria V. Zhernakova1,3, Vitaly I. Korchagin1, Andrey A. Vergun1,4, Anastasiya E. Girnyk1, Victoria A. Dikaya2,5, Marine S. Arakelyan6, Aleksey S. Komissarov1,2 and Alexey P. Ryskov1

**Abstract**

**Objectives** This study is performed in the frame of a bigger study dedicated to genomics and transcriptomics of parthenogenesis in vertebrates. Among vertebrates, obligate parthenogenesis was first described in the lizards of the genus *Darevskia*. In this genus, all found parthenogenetic species originated via interspecific hybridization. It remains unknown which genetic or genomic factors play a key role in the generation of parthenogenetic organisms. Comparative genomic and transcriptomic analysis of parthenogens and their parental species may elucidate this problem. *Darevskia valentini* is a paternal species for four (of seven) parthenogens of this genus, which we promote as a particularly important species for the generation of parthenogenetic forms.

**Data description** Total cellular RNA was isolated from kidney and liver tissues using the standard Trizol Tissue RNA Extraction protocol. Sequencing of transcriptome libraries prepared by random fragmentation of cDNA samples was performed on an Illumina HiSeq2500. Obtained raw sequences contained 117,6 million reads with the GC content of 47%. After preprocessing, raw data was assembled by Trinity and produced 491,482 contigs.

**Keywords** Caucasian rock lizards, genus *Darevskia*, interspecific hybridization, parthenogenesis, *Darevskia valentini*, transcriptome assembly

**Objective** Hybrid speciation can be considered one of the main variants of reticulate evolution [1]. In some cases, this phenomenon results in the formation of clonal lineages and parthenogenetic species. This study is performed in the frame of a bigger study dedicated to genomics and transcriptomics of parthenogenesis in vertebrates. Till now we carried out for the first time whole-genome sequencing and assembly of trio lizard species, parthenogenetic *Darevskia unisexualis*, and its parental species *D. valentini* and *D. raddei*. However, these data were not published because genome annotations were not yet done. Among vertebrates, obligate parthenogenesis was first described in the rock lizards of the genus *Darevskia* [2], which include 29 bisexual and seven unisexual (parthenogenetic) species, distributed...
in the Caucasus region, Turkey, and Iran [3]. In this genus, as in most known instances, all found parthenogenetic species originated via interspecific hybridization between closely related bisexual species [4]. Distinctive features of the Darevskia rock lizards are the high diversity of parthenogens (seven diploid forms) and ongoing hybridization events in sympatric zones of sexual and parthenogenetic species resulting in triploid and tetraploid hybrids which are considered an intermediate stage of reticulate evolution [5]. The origin of Darevskia parthenogens is phylogenetically constrained [6]. Only four parental bisexual species are involved in the origin of seven parthenogens: D. valentini and D. portschinskii as the paternal species and D. raddei and D. mixta as the maternal species [6,7]. It remains unknown, which genetic or genomic factors play a key role in the generation of clonally reproduced parthenogenetic organisms. Comparative genomic and transcriptomic analysis of parthenogens and their parental species may elucidate this problem. In particular, Darevskia valentini is a paternal species for four (of seven) Darevskia parthenogens, that we promote as a particularly important species for the generation of parthenogenetic forms.

### Data description

Samples of D. valentini for transcriptome analysis were collected in Armenia in 2016, outside of the protected areas. All individuals were hand-caught. A single adult lizard of male D. valentini from the gorge population near the Sepasar village (41°01’39.2”N, 43°48’58.0”E) was used to surgically extract organs (liver, kidneys). Before dissecting the organs, the animals were subjected to chloroform euthanasia followed by decapitation. All tissue samples were stored in RNAlater® reagent at −20 °C according to the manufacturer’s recommended protocol (Qiagen Inc.) until they were shipped to Macrogen Inc. (Korea) for RNA extraction and further transcriptome sequencing.

Total RNA was isolated from an organs/tissues using standard Trizol Tissue RNA Extraction protocol and was used to prepare the cDNA library. The paired-end sequencing libraries were prepared by random fragmentation of the cDNA samples into 350–500 bp fragments, followed by 5’ and 3’ adapter ligation using TruSeq RNA Sample Prep Kit v2 (Illumina Inc.) according to TruSeq RNA Sample Preparation Guide (Version 2, Part #15,026,495 Rev.F). Sequencing of transcriptome libraries was performed on Illumina HiSeq2500 with a mean read length of 101 bp. The Illumina Hiseq generated raw sequencing data utilizing HiSeq Control Software v2.2 for system control and base calling through an integrated primary analysis software. The BCL (base calls) binaries were converted into FASTQ format by the Illumina package bcl2fastq v1.8.4 [8] (RRID:SCR_015058). Raw transcriptome data were trimmed by Trimmomatic v0.39 to remove adapters and deduplicated by the rmdup tool [9, 10] (Data set 1) [11]. Filtered reads quality was assembled using Trinity v2.1.1 [12] with the default minimum contig length value and k-mer size parameters of 200 and 25, respectively. Summary statistics of raw samples, reads, and assembly can be accessed in Data file 1 [13]. The assembly contained 491,482 contigs with a median contig length of 923 bp (Data file 2) [14].

The annotation was provided using TransPi v1.1.0-rc pipeline [15] with OnlyAnn (only annotation) mode [16]. This option included such instruments as TransDecoder, and Trinotate. The TransDecoder program was used to predict translated proteins (Data file 3) [17]. EggNog v2.0.1 [18] was used to cross protein sequences with the Gene Ontology database. BLASTp, PFAM, and EggNOG searching tools revealed 26,812, 6496, and 15,399 proteins respectively (Data file 4) [19]. The most significant Gene Ontology terms were identified and visualized by Trinotate (Data file 5) [20].

### Table 1. Overview of data files/datasets

| Label            | Name of data file/dataset | File types (file extension) | Data repository and identifier (DOI or accession number) |
|------------------|----------------------------|-----------------------------|---------------------------------------------------------|
| Data file 1      | Summary of raw RNA and assembly characteristics | Microsoft Word file (.docx) | figshare:https://doi.org/10.6084/m9.figshare.17662030 [13] |
| Data file 2      | De novo assembly by Trinity | Fasta file (.fasta)          | NCBI Transcriptome Shotgun Assembly Sequence Data-base:https://identifiers.org/nucleotide:GJZU00000001 [14] |
| Data file 3      | TransDecoder peptides      | Peptide file (.pep)         | figshare:https://doi.org/10.6084/m9.figshare.17696930 [17] |
| Data file 4      | EggNog proteins            | Table (.csv)                | figshare:https://doi.org/10.6084/m9.figshare.17696915.v2 [19] |
| Data file 5      | Top GO terms               | Compressed PDF files (.zip) | figshare:https://doi.org/10.6084/m9.figshare.17696939 [20] |
| Data file 6      | Summary of assembled species | Compressed text, Excel and pdf files (.txt, .xls, .pdf) | figshare:https://doi.org/10.6084/m9.figshare.17696951 [21] |

The most significant GO terms were: biological process (55%); molecular function (25%) and cellular component (20%) according to biological processes ont., molecular functions ont. and cellular components ont., respectively.
of RNA polymerase II was the most over-represented category in biological processes. In molecular functions, the prevailed number of enriched genes was related to the metal ion and ATP binding. The data of top blasted species and full statistics of GO, ORF prediction numbers, and Trinotate full annotation was also performed (Data file 6) [21].

Limitations
While our transcriptome data can be used for annotation or verification of protein-coding genes in the lizard genome of D. valentini and related lizard species, some limitations are connected with a restricted number of tissues (only liver and kidney) taken for generation of the mixed transcriptome.

Abbreviations
cDNA complementary deoxyribonucleic acid
RNA ribonucleic acid
BCL binary base calls
bp base pair
GO Gene Ontology
ORF open reading frame
ATP adenosine triphosphate

Acknowledgements
RNA characterization experiments were performed using the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, IGB RAS.

Authors’ contributions
SR, DZ, VO performed the assembly, analysis, and interpretation of the raw sequence data. VK, AG, AV designed the sampling methods. SR, AV, AR wrote the manuscript. MA collected the samples. AR, AK designed the study. All authors have read and approved the final manuscript.

Funding
This research was funded by the Russian Science Foundation (RSF) Research Project № 19-14-00083.

Availability of data and materials
The raw data described in this Data note can be freely and openly accessed on the NCBI SRA database under accession ID SRX14421363. Please see Table 1 for details and links to the rest of the data [11, 13, 14, 17, 19–21].

Declarations
Ethics approval and consent to participate
The study was approved by the Ethics Committee of the Moscow State University (Permit Number: 24–01) and conducted strictly according to ethical principles and scientific standards. Alive-animal handling procedures were approved by Yerevan State University according to the ethical guidelines, capture permit Code 5/22.1/51043 was issued by the Ministry of Nature Protection of the Republic of Armenia for scientific studies.

Consent for publication
Not applicable.

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

Received: 28 May 2022 / Accepted: 14 October 2022
Published online: 08 November 2022

References
1. Datsyshansky T. Genetics and the origin of species. New York: Columbia Univ. Press; 1937.
2. Darevskii IS. Rock lizards of the Caucasus: systematics, ecology, and phylogeny of the polymorphic groups of Caucasian rock lizards of the subgenus Aracneolacerta. Nauka. 1967;1–216.
3. Uetar P, Freed P, Holsek J, et al. THE REPTILE DATABASE: http://www.reptile-database.org/. Accessed 3 Apr 2021.
4. Neaves WR, Baumann UB. Unidirectional reproduction among vertebrates. Trends Genet. 2011;27:81–8. doi:https://doi.org/10.1016/j.tig.2010.12.002.
5. Danielbay F, Atyekly M, Stepanian I. The progress of microevolution in hybrids of rock lizards of genus Darevskia. Biol J Linnean Soc. 2008;60:147–56.
6. Murphy RW, Fu J, Macculloch RD, Darevskiy IS, Kupriyanova LA. A fine line between sex and unisexuals: The phylogenetic constraints on parthenogenesis in lacertid lizards. Zool J Linnean Soc. 2000;130:527–49.
7. Fu J, Murphy RW, Darevskiy IS. Toward the phylogeny of caucasian rock lizards: implications from mitochondrial DNA gene sequences (Reptilia: Lacertidae). Zool J Linnean Soc. 1997;120:463–77. doi:https://doi.org/10.1111/j.1096-3642.1997.tb00128.x.
8. bcftools Conversion Software. https://support illumina.com/sequencing/ sequencing_software/bcftools-conversion-software.html. Accessed 11 Apr 2022.
9. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. doi:https://doi.org/10.1093/bioinformatics/btu170.
10. aglavs/mdup. Removes optical duplicates from raw Illumina sequence reads. GitHub. (n.d.). https://github.com/aglavs/mdup. Accessed 11 Apr 2022.
11. RNA-seq of D.valentini. adult male mixed liver and kidneys. NCBI Sequence Read Archive https://identifiers.org/ncbi/insdc.sra:SRX14421363 (2022).
12. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data with a reference genome. Nat Biotechnol. 2011;29:644–52. doi:https://doi.org/10.1038/nbt.1883.
13. Ryakhovsky S. Summary of raw RNA and assembly characteristics. figshare. 2022. https://doi.org/10.6084/m9.figshare.17762030.
14. Ryakhovsky S. De novo assembly by Trinity. NCBI Transcriptome Shotgun Assembly Sequence Database https://identifiers.org/nucleotide-GZ/1000000001 (2022).
15. Rivera-Vicéns RE, Garcia-Escudero CA, Conci N, Etel M, Wörheide G. TransPi – a comprehensive TRanscriptome ANalysis PIpeline for de novo transcriptome assembly. bioRxiv. 2021;2021.02.18.431773. doi:https://doi.org/10.1101/2021.02.18.431773.
16. Ryakhovsky SS, Dikaya VA, Khorchin VI, Vergun AA, Danilov LG, Ochkalov SD, et al. De novo transcriptome assembly and annotation of parthenogenetic lizard Darevskia unisexualis and its parental ancestors Darevskia valentini and Darevskia raddei nainens. Data Br. 2021;39:107685.
17. Ryakhovsky S. TransDecoder. peptides. figshare. https://doi.org/10.6084/m9.figshare.17696930 (2022).
18. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Bio Evol. 2021;38:5825–9. doi:https://doi.org/10.1093/molbev/msab293.
19. Ryakhovsky S, BLASTp PFAM, EggNOG. proteins. figshare https://doi.org/10.6084/m9.figshare.17696915.v2 (2022).
20. Ryakhovsky S. Top GO terms. figshare https://doi.org/10.6084/m9.figshare.17696939 (2022).
21. Ryakhovsky S. Summary of Trinotate, TransDecoder and top blasted species. figshare. https://doi.org/10.6084/m9.figshare.17696951 (2022).