Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato

Marko Petek1,6, Maja Zagorščak1,6, Živa Ramšak1, Sheri Sanders3, Špela Tomaž1,2, Elizabeth Tseng4, Mohamed Zouine5, Anna Coll1 & Kristina Gruen1

Although the reference genome of Solanum tuberosum Group Phureja double-monoploid (DM) clone is available, knowledge on the genetic diversity of the highly heterozygous tetraploid Group Tuberosum, representing most cultivated varieties, remains largely unexplored. This lack of knowledge hinders further progress in potato research. In conducted investigation, we first merged and manually curated the two existing partially-overlapping DM genome-based gene models, creating a union of genes in Phureja scaffold. Next, we compiled available and newly generated RNA-Seq datasets (cca. 1.5 billion reads) for three tetraploid potato genotypes (cultivar Désirée, cultivar Rywal, and breeding clone PW363) with diverse breeding pedigrees. Short-read transcriptomes were assembled using several de novo assemblers under different settings to test for optimal outcome. For cultivar Rywal, PacBio Iso-Seq full-length transcriptome sequencing was also performed. EvidentialGene redundancy-reducing pipeline complemented with in-house developed scripts was employed to produce accurate and complete cultivar-specific transcriptomes, as well as to attain the pan-transcriptome. The generated transcriptomes and pan-transcriptome represent a valuable resource for potato gene variability exploration, high-throughput omics analyses, and breeding programmes.

Background & Summary
At species level, genomes of individuals can differ in single nucleotide polymorphisms (SNPs), short insertions and deletions (INDELs), gene copy numbers, and presence or absence of genes1. The latter leads to the concept of species specific pan-genomes, namely the core genome present in most individuals and the dispensable genome comprised of genes present only in a subset of individuals, which results in the emergence of particular subgroup-specific phenotypes. This concept has been extended to pan-transcriptomes, where the presence or absence of variations is not bound only to gene content, but also to genetic and epigenetic regulatory elements. Pan-genomes and pan-transcriptomes have been described in the model plant species Arabidopsis thaliana2 and several crop species, including maize3,4, rice5, wheat6 and soybean7.

While the genome of a double-monoploid clone of Solanum tuberosum Group Phureja (DM) is available8, this diploid potato group differs from the tetraploid Group Tuberosum, which includes most varieties of cultivated potato. Through domestication and modern breeding efforts, different potato cultivars also acquired genes from other closely related Solanum species or lost some ancestral genes1. Different breeding programmes have resulted in accumulation of smaller genome modifications, e.g. SNPs and INDELs. Consequently, each distinct potato cultivar harbours a unique set of transcripts, resulting in physiological and developmental differences and different responses to biotic and abiotic stress. SNP and INDEL profile differences and novel gene variants in anthocyanin pathway were identified in a comparative transcriptome analysis of two Chinese potato cultivars9. Unfortunately, we could not include these transcriptomes in our pan-transcriptome because the assemblies were not publicly accessible.

Based on the DM genome, the PGSC and ITAG annotation consortia8,10 have each independently produced potato gene models. For practical reasons, most potato researchers use only one genome annotation, either PGSC or ITAG, especially when conducting high-throughput analyses. Using an incomplete gene set can lead to false
conclusions regarding gene presence or gene family diversity in potato. Using a computational pipeline followed by manual curation, we have consolidated the two publicly available Group Phureja DM gene model sets to produce a unified one.

While a combined DM gene set is useful, it is still not as useful as a pan-transcriptome that includes assemblies from cultivated potatoes. However, obtaining an optimal transcriptome from short-read RNA-Seq data is not a trivial task. Each de novo assembler suffers from different intrinsic error generation and no single assembler performs best on all datasets. To maximise diversity and completeness of potato cultivar transcriptomes, usage of multiple de novo transcriptome assemblers and various parameter combinations over the same input data was employed. Following this “over-assembly” step, we used the tr2aadcs pipeline from EvidentialGene to reduce redundancy across assemblies and obtain cultivar-specific transcriptomes. Finally, we consolidated representative cultivar-specific sequences to generate a potato pan-transcriptome (StPanTr). These transcriptomes will improve high throughput sequencing analyses, from RNA-Seq and sRNA-Seq to more specific ones like ATAC-Seq, by providing a more comprehensive and accurate mapping reference. The translated protein sequences can enhance the sensitivity of high-throughput mass-spectroscopy based proteomics. The resource is valuable also for the design of various PCR assays, e.g. quantitative PCR, where exact sequence information is required. Additionally, the knowledge generated regarding variations in transcript sequences between cultivars, such as SNPs, insertions and deletions, will be a key instrument in assisting potato breeding programmes.

Methods

Merging PGSC and ITAG gene models of reference genome Group Phureja. GFF files corresponding to their respective gene models (PGSC v4.04, ITAG v1.0) were retrieved from the Spud DB (solanaceae.plantbiology.msu.edu) potato genomics resource. The two models (39,431 PGSC and 35,004 ITAG) were then compared on the basis of their exact chromosomal location and orientation in order to create the most complete set of genes from both predicted combination models. Several combinations arose from the merge (Fig. 1a), those for which no action was required (singletons, i.e. sole PGSC or ITAG genes not covering any other genes); 1-to-1 or 1-to-2 combinations between PGSC and ITAG genes, which were solved programmatically; and lastly, combinations of more than 3 genes in various combination types, which continued on to manual curation. The latter (318 gene clusters; example in Fig. 1b) were considered to be non trivial merge examples (overlapping genes in two models or multiple genes in PGSC corresponding to a single gene in ITAG). This resulted in a merged DM genome GFF3 file with 49,322 chromosome position specific sequences, of which 31,442 were assigned with ITAG gene IDs and 17,880 with PGSC gene IDs.

Data pre-processing. The complete bioinformatic pipeline is outlined in Fig. 2. Sequence quality assessment of raw RNA-Seq data, quality trimming, and removal of adapter sequences and polyA tails was performed using CLC Genomics Workbench v6.5-v10.0.1 (Qiagen) with maximum error probability threshold set to 0.01 (Phred quality score 20) and no ambiguous nucleotides allowed. Minimal trimmed sequence length allowed was set to 15 bp while maximum up to 1 kb. Orphaned reads were re-assigned as single-end (SE) reads. Processed reads were pooled into cultivar datasets as properly paired-end (PE) reads or SE reads per cultivar per sequencing platform. For the Velvet assembler, SOLiD reads were converted into double encoding reads using perl script "denovo_preprocessor_solid_v2.2.1.pl". To reduce the size of cv. Désirée and cv. Rywal datasets, digital normalization was performed using khmer from bbmap suite v37.68 prior to conducting de novo assembly using Velvet and rnaSPAdes.

PacBio long reads were processed for each sample independently using Iso-Seq 3 analysis software (Pacific Biosciences). Briefly, the pipeline included Circular Consensus Sequence (CCS) generation, full-length reads identification (“classify” step), clustering isoforms (“cluster” step) and “polishing” step using Arrow consensus algorithm. Only high-quality full-length PacBio isoforms were used as input for further steps.

PacBio Cupcake ToFU pipeline. Cupcake ToFU (github.com/Magdoll/cDNA_Cupcake) scripts were used to further refine the Iso-Seq transcript set. Redundant PacBio isoforms were collapsed with “collapse_isofoms_by_sam.py” and counts were obtained with “get_abundance_post_collapse.py”. Isoforms with less than two supporting counts were filtered using “filter_by_count.py” and 5’-degraded isoforms were filtered using “filter_away_subset.py”. Isoforms from the two samples were combined into one non-redundant Iso-Seq transcript set using “chain_samples.py”.

De Bruijn graph based de novo assembly of short reads. Short reads were de novo assembled using Trinity v.2013-02-25, Velvet/Oases v. 1.2.10, rnaSPAdes v.3.11.1 and CLC Genomics Workbench v8.5.4-v10.1.1 (Qiagen). Illumina and SOLiD reads were assembled separately. For CLC Genomics de novo assemblies, combinations of three bubble sizes and 14 k-mer sizes were tested on PW363 Illumina dataset. Varying bubble size length did not influence the assembly statistics much (Supplementary Fig. 2), therefore we decided to use the length of 85 bp for Illumina datasets of the other two cultivars. Bubble size and k-mer length parameters used for Velvet and CLC are given in Table 1. The scaffolding option in CLC and Velvet was disabled. More detailed information per assembly is provided in Auxiliary Table 2.

Decreasing redundancy of assemblies and annotation. 739 mio Désirée short reads were assembled into 3,765,661 potential transcripts, 284 mio PW363 short reads were assembled into 6,022,291 potential transcripts, and 710 mio Rywal short reads and 1.4 mio Rywal PacBio sequences were assembled into 1,912,821 potential transcripts. While generation of several transcriptomes from diverse input data and various parameter combinations increases the likelihood of capturing and accurately assembling transcripts, redundancy reduction without loss of information and error removal from the over-assemblies is required. All cultivar-specific transcriptome assemblies, compiled into cultivar-specific transcriptome over-assembly, were initially filtered with the tr2aadcs pipeline (part...
of EvidentialGene v2016.07.11) which consists of four steps. First, all perfect redundant nucleotide sequences are removed using fastnrdb, part of the exonerate package, leaving only the transcript with the longest coding region. Next, all perfect fragments of the remaining transcripts are removed using cd-hit. These first two steps are important in reducing transcriptome redundancy, as true transcripts are expected to be assembled independently by multiple of the assembly methods. Keeping the transcripts with the longest and most complete coding region helps eliminate chimeric and misassembled transcripts, as these errors tend to occur more often in UTR regions or in a manner that causes frameshifts and long, incomplete coding regions.

The third and the fourth step of the tr2aacds pipeline segregate transcripts that are likely isoforms, alleles, or other variations seen at a single locus. This is done through amino acid sequence clustering, which identifies putative transcripts that vary only in silent mutations, and through reciprocal BLAST, which detects high-identity exon-sized alignments (likely isoforms). A tag is assigned to all transcripts providing detailed information on why they were discarded (e.g. perfect fragments, perfect duplicates, very high similarity, …) or why they were marked.
as alternatives (and which sequence they are an alternative form of). The final output of the tr2aacds pipeline are three sets of data – a non-redundant set of representative sequences (i.e. main set), a set of putative alternatives mapped to the representative set (i.e. alt set), and a discarded set (i.e. drop set) of redundant sequences. It is important to note that not all dropped sequences are of poor quality or incorrect – many of them are dropped due to full or partial redundancy.

Representative and alternative sets (termed okay sets by EvidentialGene) were merged into initial cultivar reference transcriptomes and, as tr2aacds only uses internal evidence for data curation, used in further external evidence for assembly validation, filtering and annotation steps (Fig. 2). The de novo cultivar-specific transcript sets were first mapped to the DM reference genome by STARlong 2.6.1d using parameters optimized for de novo transcriptome datasets (all scripts are deposited at FAIRDOMHub (fairdomhub.org) project home page).

Aligned transcripts were analysed with MatchAnnot to identify transcripts that match the PGSC or ITAG gene models. Domains were assigned to the polypeptide dataset using InterProScan software package v5.37-71.0.

For all transcripts and coding sequences, annotations using DIAMOND v0.9.24.125 were generated by querying UniProt (www.uniprot.org) retrieved databases (E-value cut-off 10 \(^{-0.5}\) and query transcript/cds and target sequence alignment coverage higher or equal to 50%; custom databases: Solanum tuberosum, Solanaceae, plants). Initially assembled transcriptomes were also screened for nucleic acid sequences that may be of vector origin (vector segment contamination) using VecScreen plus taxonomy program v0.16 against NCBI UniVec Database (ftp.ncbi.nlm.nih.gov/pub/UniVec).

Potential biological and artificial contamination was identified as up to 3.3% of sequences per cultivar, if taking into account cases when potential contaminants covered less than 1% of the sequence (number of sequences with strong, moderate and weak proof of contamination as follows: 182, 547 and 10,509 for Désirée; 48, 228 and 7,877 for PW363; 169, 179 and 4,103 for Rywal). The results from
MatchAnnot, InterProScan and DIAMOND were used as biological evidence in further filtering by in-house R scripts. Transcripts that did not map to the genome nor had any significant hits in either InterPro (www.ebi.ac.uk/interpro) or UniProt (www.uniprot.org) were eliminated from further analysis to obtain higher reliability of constructed transcriptomes. Pajek v5.08, in-house scripts, and cdhit-2d from the CD-HIT package v4.6.2 were used to re-assign post-filtering representative and alternative classes and to obtain finalised cultivar-specific transcriptomes.

The whole redundancy removal procedure reduced the initial transcriptome assemblies by 18-fold for Désirée, 38-fold for Rywal, and 24-fold for PW363. Completeness of each initial de novo assembly to cultivar-specific transcriptome was estimated with BUSCO (Supplementary Figs. 1–3).

Individual contributions by various assembly methods were investigated in light of what contributed to the final, cleaned cultivar transcriptomes. SOLiD assemblies (Supplementary Fig. 1: CLCdnDe1, CLCdnDe8, VdnDe8–10), produced by either CLC or Velvet/Oases pipelines, contributed least to transcriptomes, which can mostly be attributed to short length of the input sequences. Interestingly, increasing k-mer size in the CLC pipeline for Illumina assemblies produced more complete assemblies according to BUSCO scores and more transcripts were selected for the initial transcriptome (Supplementary Fig. 1: CLCdnDe1–7, CLCdnDe9–14). On the contrary, increasing k-mer length in Velvet/Oases pipeline lead to transcripts that were less favoured by the redundancy removal procedure (Supplementary Fig. 1: VdnDe1–7). The Trinity assembly was comparable

| Genotype | Assembly ID | Read type | Assembler | Assembler version | k-mer length (word size) | Bubble size |
|----------|-------------|-----------|-----------|------------------|--------------------------|-------------|
| Désirée  | CLCdnDe8    | SOLiD     | CLC de novo | 9.1              | 24                       | 50          |
| Désirée  | CLCdnDe1    | SOLiD     | CLC de novo | 10.0.1           | 24                       | 50          |
| Désirée  | VdnDe8, …   | SOLiD     | Velvet/Oases | 1.2.10        | 23, 33, 43               | Default     |
| Désirée  | CLCdnDe9, … | Illumina  | CLC de novo | 9.1              | 21, 23, 33, 43           | 85          |
| Désirée  | …, CLCdnDe14|          |            |                  |                          |             |
| Désirée  | CLCdnDe2, … | Illumina  | CLC de novo | 10.0.1           | 21, 23, 33, 43           | 85          |
| Désirée  | …, CLCdnDe7 |          |            |                  |                          |             |
| Désirée  | TDe         | Illumina  | Trinity    | 2013-02-25       | 25                       | NA          |
| Désirée  | VdnDe1, …   | Illumina  | Velvet/Oases | 1.2.10        | 23, 33, 43, 53           | Default     |
| Désirée  | …, VdnDe7   |          |            | 63, 73, 83       |                          |             |
| PW363    | CLCdnPW1    | SOLiD     | CLC de novo | 8.5.4            | 24                       | 50          |
| PW363    | CLCdnPW2    | SOLiD     | CLC de novo | 9.1              | 24                       | 50          |
| PW363    | VdnPW8, …   | SOLiD     | Velvet/Oases | 1.2.10        | 23, 33, 43               | Default     |
| PW363    | …, VdnPW10  |          |            | 63, 73, 83       |                          |             |
| PW363    | CLCdnPW3, … | Illumina  | CLC de novo | 8.5.4            | 21, 23, 24, 25, 30, 50, 65, 85 |
| PW363    | …, CLCdnPW44|          |            |                  |                          |             |
| PW363    | CLCdnPW45, …| Illumina  | CLC de novo | 10.0.1           | 21, 23, 33, 43, 85       |
| PW363    | …, CLCdnPW50|          |            |                  |                          |             |
| PW363    | SdnPW1      | illumina  | rnaSPAdes   | 3.11.1           | 43                       | Default     |
| PW363    | TPW         | illumina  | Trinity     | 2013-02-25       | 25                       | NA          |
| PW363    | VdnPW1, …   | illumina  | Velvet/Oases | 1.2.10        | 23, 33, 43, 53           | Default     |
| PW363    | …, VdnPW7   |          |            | 63, 73, 83       |                          |             |
| Rywal    | PRdnRY1     | PacBio    | Iso-Seq. 3, | 2017            | NAa                      | NAa         |
| Rywal    | …, Cupcake  | ToFU     |            |                  |                          |             |
| Rywal    | CLCdnRY1, … | illumina  | CLC de novo | 9.1              | 21, 23, 33, 43           | 85          |
| Rywal    | …, CLCdnRY6 |          |            |                  |                          |             |
| Rywal    | CLCdnRY7, … | illumina  | CLC de novo | 10.1.1           | 21, 23, 33, 43           | 85          |
| Rywal    | …, CLCdnRY12|          |            |                  |                          |             |
| Rywal    | SdnRY1      | illumina  | rnaSPAdes   | 3.11.1           | 43                       | Default     |
| Rywal    | VdnRY1, …   | illumina  | Velvet/Oases | 1.2.10        | 23, 33, 43, 53           | Default     |
| Rywal    | …, VdnRY7   |          |            | 63, 73, 83       |                          |             |

Table 1. Parameters used for short read de novo assembly generation. NAp – not applicable. NA – not available.
to the high k-mer CLC assemblies in transcriptome contribution and BUSCO score (Supplementary Fig. 1). It might seem that the PacBio Iso-Seq transcripts contributed less than expected to the cv. Rywal transcriptome (Supplementary Fig. 3), however it should be noted that a considerable number of PacBio transcripts was assigned to the EvidentialGene drop set because they had perfect or near-perfect CDS identity of transcripts assembled by CLC. The EvidentialGene pipeline also prioritised CLC-assembled transcripts over PacBio transcripts because the redundancy removal algorithm reorders the near-perfect duplicates by transcript name and only retains the first transcript listed (Auxiliary Table 634).

**Potato pan-transcriptome construction.** While the PGSC gene model defined transcripts as well as coding sequences, the ITAG gene model defined only coding sequences. Therefore, the potato pan-transcriptome construction was conducted at the level of CDS.

Cultivar-specific representative coding sequences (57,943 of Désirée, 43,883 of PW363 and 36,336 of Rywal) were combined with coding sequences from the merged Phureja DM gene models (17,880 and 31,442 non-redundant PGSC and ITAG genes, respectively) and subjected to the cdhit-est 24 algorithm (global sequence identity threshold 90%, alignment coverage for the shorter sequence 75%, bandwidth of alignment 51 nt and word length of 9) to create potato pan-trancriptome. Sequences that did not cluster using cdhit-est were separated into tetraploid and DM datasets and subjected to the cdhit-2d 24 algorithm (local sequence identity threshold 90%, alignment coverage for the shorter sequence 45%, bandwidth of alignment 45 nt and word length of 5).

Sequences that are shared by the DM merged gene model and de novo assembled cultivar-specific transcriptomes were designated as "core" transcripts, and sequences that were assembled in only one transcriptome were designated "genotype-specific". The total pan-transcriptome includes 96,886 representative, non-redundant transcripts and 90,618 alternative sequences (covering alternative splice forms, allelic isoforms and partial transcripts) for those loci (Fig. 3, Supplementary Fig. 4, Auxiliary Table 735). The core subset of the pan-transcriptome contains 68,708 sequences, among which 12% are partial sequences.

Polyploid crop pan-genomes generally consist of many cultivar-specific genes 36. Therefore we included all genotype-specific sequences in our potato pan-transcriptome (Fig. 3, Supplementary Fig. 4). This subset contains 64,529 sequences, among which 13% sequences are partial 35. Genotype-specific transcripts are generally shorter in length than the core transcripts, however they do not differ much in the percentage of complete transcripts.

![Fig. 3](https://example.com/fig3.png)
| Genotype     | Sample descriptiona | Sequencing platform | Library typeb | Number of readsd | SRA ID                  |
|--------------|---------------------|---------------------|---------------|------------------|------------------------|
| Désirée      | PVY inoculated leaves | Illumina            | DSN-normalized | ~54 mio          | SRR10070125            |
|              |                     |                     | PE90 unstranded |                 |                        |
| Désirée      | non-transformed and PVY- inoculated plants, non-infested and CPB infested leaves | Illumina            | PE90 unstranded | ~195 mio          | SRR1207287, …, SRR1207290 |
| Désirée      | mock and PVY inoculated leaves and stem | SOLiD              | SE50 unstranded | ~154 mio          | SRR10065428, SRR10065429 |
| Désirée      | leaves               | Illumina            | SE50 unstranded | ~172 mio          | SRR3161991, SRR3161995, SRR3161999, SRR3162003, SRR3162007, SRR3162011, SRR3162015, SRR3162019, SRR3162023, SRR3162027, SRR3162031, SRR3162035 |
| Désirée      | seedlings            | Illumina            | SE100 unstranded | ~80 mio          | SRR4125238, …, SRR4125247 |
| Désirée      | roots                | Illumina            | SE100 unstranded | ~31 mio          | SRR4125248, …, SRR4125252 |
| Désirée      | mock and Phytophthora infestans inoculated leaves | Illumina            | PE90 unstranded | ~53 mio          | ERR305632               |
| Rywal        | mock and PVY inoculated leaves | PacBio              | Iso-Seq 0.7–2 Kb, 2–3.5 Kb, >3.5 Kb | ~1.4 mio | SRR8281993, …, SRR8282008 |
|             |                     |                     | CCS           |                  |                        |
| Rywal        | mock and PVY inoculated leaves | Illumina            | PE100 strand-specific | ~710 mio | SRX6801457, …, SRX6801468 |
| PW363        | PVY inoculated leaves | Illumina            | DSN-normalized | ~104 mio          | SRR10070123, SRR10070124 |
| PW363        | mock and PVY inoculated leaves | SOLiD              | SE50 unstranded | ~180 mio          | SRR10065430, …, SRR1006543 |

Table 2. Samples used to generate the de novo transcriptome assemblies. aPVY, <em>Phytophthora infestans</em> inoculated. bSE, single-end library (the number stands for read length in nt); DSN-normalized, RNA-Seq library utilizing the crab duplex nuclease; CCS, circular consensus sequences. dFor paired-end libraries, pairs are counted as two reads.

### Data records

Transcriptomic sequences of three potato genotypes, cv. Désirée, cv. Rywal and breeding clone PW363, were obtained from in-house RNA-Seq projects32–41 and supplemented by publicly available cv. Désirée datasets retrieved from SRA42–44 (Table 2). All generated files have been deposited to FAIRDOMHub45 under project name _p_atRT (https://fairdomhub.org/projects/161). The largest quantity of reads, cca. 739 mio reads of various lengths, was obtained for cv. Désirée, using Illumina and SOLiD short-read sequencing platforms. For cv. Rywal and breeding clone PW363 only mature leaf samples were available. For cv. Désirée leaf samples were augmented with samples from stems, seedlings and roots. For cv. Rywal short-read sequencing was complemented with full-length PacBio Iso-Seq sequencing of independent samples. Detailed sample information is provided in Auxiliary Table 145.

Cv. Rywal NPR1-1 coding sequences, sequenced by the Sanger method, were deposited at NCBI GenBank under accession numbers MT210578, …, MT210585.46–53. The GTF file with merged ITAG and PGSC gene models for <em>S. tuberosum</em> Group Phureja DM genome v4.044 is also available at FAIRDOMHub project page, as well as the cultivar-specific and pan-transcriptome assembly FASTA and annotation files44.

### Technical Validation

As a measure of assembly accuracy, the percentage of correctly assembled bases was obtained by mapping Illumina reads back to cultivar-specific initial transcripts using STAR v.2.6.1d RNA-seq aligner with default parameters (Table 3). To assess the quality of transcriptomes via size-based and reference-based metrics, we ran TransRate v 1.0.155 on cultivar-specific transcriptomes (prior to and post filtering, Table 4), cultivar-specific representative transcript sequences and PGSC defined representative transcripts (Table 5). Comparative metrics for cultivar-specific coding sequences (CDS) were obtained using Conditional Reciprocal Best BLAST (CRBB)46 against merged Phureja DM gene model coding sequences.

To estimate the measure of completeness and define the duplicated fraction of assembled transcriptomes (prior and post filtering cultivar-specific transcriptomes and pan-transcriptome), BUSCO v3.97 scores were calculated using embryophyta_odb958 lineage data (Table 6). At the cultivar-specific transcriptome level, the most diverse dataset in terms of tissues and experimental conditions, the cv. Désirée dataset, resulted in the highest BUSCO score as expected. The success in classification of representative and alternative transcripts is evident from the pan-transcriptome BUSCO scores (i.e. differences in single-copy and duplicated BUSCOs for representative and alternative dataset). The highest number of fragmented BUSCOs is observed for the breeding clone PW363, which can probably attribute to the highest number of short-contig assemblies. Furthermore, the long-read assembly presumably contributed to the shift in favour of single-copy BUSCOs (Table 6) and uniquely mapped reads (Table 3) for cv. Rywal.

To inspect the quality of paralogue cluster assignments, multiple sequence alignments (MSA) using MAFFT v7.27159 were conducted on representative and alternative sequences from paralogue clusters50 containing sequences from each of the four genotypes (Désirée, PW363, Rywal and DM). Alignments were visualized using MView v1.6650 (Auxiliary File 264). These alignments were used to quality check de novo transcriptome assemblies.
and helped us optimise the pipeline. The alignments within groups showed differences that can be attributed to biological diversity, e.g. SNPs and INDELS as well as alternative splicing61.

To estimate the proportion of transcripts originating from organisms other than potato, we performed a taxonomic classification of all cultivar-specific transcriptome sequences using Centrifuge v1.0.462 and the NCBI nt database. We used pavian63 to generate classification summary reports (Auxiliary File 3 64). The transcriptomes include altogether less than 1% bacterial, viral, fungal and protozoan transcripts. Other non-plant sequences are from common plant pests such the arachnid \textit{Tetranychus urticae} in PW363 transcriptome (cca. 4%) and insects \textit{Leptinotarsa decemlineata} (cca. 0.1%) and \textit{Myzus persicae} (cca. 0.1%) in the Désirée transcriptome64.

Sanger sequencing confirmation of assembled transcripts. To further validate the quality of the constructed Rywal reference transcriptome, the cultivar-specific assembled transcript coding for NPR1-1 protein in cv. Rywal was compared to sequences amplified from isolated cDNA and sequenced by Sanger method. Total RNA was isolated from 4-week-old cv. Rywal plants using RNeasy Plant Mini kit (Qiagen). Residual genomic DNA was digested with DNase I (Qiagen). Treated total RNA was reversed transcribed using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). The full-length CDS of NPR1-1 (Sotub07g016890.1.1, ITAG genome annotation) was amplified from the cv. Rywal cDNA with forward (5′-ATGGAGAGCGGCCACGAGA-3′) and reverse (5′-CTACTTTTTTTCAAACTTTGACTGACATT-3′) primers. The PCR product was inserted into pJET1.2/blunt vector with CloneJET PCR Cloning Kit (Thermo Scientific) and introduced into OneShot TOP10 Chemically Competent \textit{Escherichia coli} cells. The plasmids were isolated from 8 transformed colonies, grown on selection media, using the GenElute Plasmid Miniprep Kit (Sigma-Aldrich). Inserts were sequenced (GATC Services, Eurofins Genomics) using pJET1.2 sequencing primers (Thermo Scientific) as well as insert-specific primers, designed to ensure full sequence coverage in both directions.

A multiple sequence alignment comparing Sanger sequences with NPR1-1 coding sequence from the assembled cv. Rywal transcriptome (parologue cluster stCuSTr-R_29366) and Phureja DM gene model (Sotub07g016890.1.1) was constructed using Geneious Prime 2020.1.1. The Sanger

| Mapping statistics/genotype           | Désirée' | PW363' | Rywal' |
|---------------------------------------|----------|--------|--------|
| Number of input reads                 | 177,149,132 | 52,171,015 | 342,767,035 |
| Average input read length             | 178      | 179    | 199    |

**UNIQUE READS:**
- Uniquely mapped reads number       | 64,507,790 | 18,416,487 | 206,003,021 |
- Uniquely mapped reads* %           | 36%       | 35%     | 60%     |
- Average mapped length              | 175       | 176     | 196     |
- Number of splices*: Total          | 496,170   | 267,268 | 1,700,235 |
- Number of splices*: Annotated (sjdb) | 0         | 0       | 0       |
- Number of splices*: GT/AG          | 258,208   | 105,551 | 1,162,885 |
- Number of splices*: GC/AG          | 10,749    | 5,693   | 79,495  |
- Number of splices*: AT/AC          | 1,486     | 2,192   | 1,840   |
- Number of splices*: Non-canonical  | 225,727   | 153,832 | 456,015 |
- Mismatch rate per base %           | 0.50%     | 0.53%   | 0.59%   |
- Deletion rate per base              | 0.03%     | 0.03%   | 0.03%   |
- Deletion average length             | 2.72      | 2.53    | 3.02    |
- Insertion rate per base             | 0.02%     | 0.02%   | 0.03%   |
- Insertion average length            | 1.93      | 1.86    | 1.91    |

**MULTI-MAPPING READS:**
- Number of reads mapped to multiple loci | 98,694,222 | 29,366,122 | 108,669,657 |
- % of reads mapped to multiple loci*   | 56%       | 56%     | 32%     |
- Number of reads mapped to too many loci | 4,652,918 | 1,555,704 | 1,541,238 |
- % of reads mapped to too many loci   | 2.63%     | 2.98%   | 0.45%   |

**UNMAPPED READS:**
- % of reads unmapped: too many mismatches | 0%       | 0%     | 0%     |
- % of reads unmapped: too short       | 5.25%     | 5.43%   | 7.75%   |
- % of reads unmapped: other           | 0%       | 0%     | 0%     |

**CHIMERIC READS:**
- Number of chimeric reads             | 0         | 0       | 0       |
- % of chimeric reads                  | 0%       | 0%     | 0%     |

Table 3. Transcriptome quality control by RNA-seq reads remapping. Illumina paired-end reads used for generating assemblies were mapped back to the corresponding cultivar specific transcriptomes using STAR. *Number of reads crossing supposed splice sites. ’Initially constructed transcriptomes (prior to filtering steps). †Relevant % of mapped reads: % of uniquely mapped reads + % of reads mapped to multiple loci.
sequencing revealed the presence of several distinct gene variants in the analysed colonies, differing in the SNP pattern. The PBdnRY1_9437 sequence is validated by complete sequence identity with the two colonies (seq. 7 and 8), while the CLCdnRY_10265 shares two SNPs with seq. 4–6, matching the SNP pattern partially. Although the Phureja DM gene model also shares two SNPs with some of the colonies, its overall SNP pattern differs significantly from cv. Rywal, distinguishing the cultivar-specific transcripts from that of Phureja DM (Fig. 4).

**Usage Notes**

**Insights into variability of potato transcriptomes.** Based on the comparison of cultivar-specific transcriptomes we identified cca. 23,000, 13,000, and 7,500 paralogue groups of transcripts in cv. Désirée, breeding clone PW363 and cv. Rywal, respectively, not present in the merged Phureja DM gene model. The addition of
Iso-Seq dataset in the case of cv. Rywal confirms that long reads contribute to less fragmentation of the de novo transcriptome. It is therefore recommended to generate at least a subset of data with one of the long-read technologies to complement the short-read RNA-seq. As it can be seen by the reduction rate in PW363 (24-fold), producing additional short-read assemblies does not contribute to the transcriptome quality as much as having several tissues or a combination of 2nd and 3rd generation sequencing (38-fold Rywal).

From all four genotypes, cv. Désirée has the highest number of cultivar-specific representative transcripts, which can be attributed to having the most diverse input dataset used for the de novo assemblies in terms of tissues sequenced (stem, seedlings and roots) and experimental conditions covered. Cv. Désirée also benefited from the inclusion of a DSN Illumina library to capture low level expressed transcripts. However, even the leaf-specific reference transcriptomes of cv. Rywal and breeding clone PW363 include thousands of specific genes, indicating that cultivar specific gene content is common. Remarkably, we identified several interesting features when inspecting paralogue groups of transcripts, demonstrating the variability of sequences in potato haplotypes and the presence of the alternative splicing variants that contribute to the pan-transcriptome (Fig. 5, Auxiliary File 261).

### Table 6. Assessment of completeness of constructed transcriptomes. Percentage of BUSCOs identified in each transcriptome assembly step. (S): Complete and single-copy BUSCOs %; (D): Complete and duplicated BUSCOs % (C): Complete BUSCOs (S + D) % (F): Fragmented BUSCOs % (M): Missing BUSCOs % rep: representative alt: alternative *

| cv. Désirée | initial rep+alt | post 1st filtering rep+alt | final rep+alt |
|-------------|----------------|---------------------------|--------------|
| (S)         | 37.8           | 37.8                      | 37.4         |
| (D)         | 59.4           | 59.2                      | 58.4         |
| (C)         | 97.2           | 97.0                      | 95.8         |
| (F)         | 1.1            | 1.2                       | 1.4          |
| (M)         | 1.7            | 1.8                       | 2.8          |
| breeding clone PW363 | initial rep+alt | post 1st filtering rep+alt | final rep+alt |
| (S)         | 39.9           | 39.2                      | 38.4         |
| (D)         | 51.7           | 51.2                      | 50.9         |
| (C)         | 91.6           | 90.4                      | 89.3         |
| (F)         | 2.9            | 3.4                       | 3.5          |
| (M)         | 5.6            | 6.2                       | 7.2          |

| cv. Rywal | initial rep+alt | post 1st filtering rep+alt | final rep+alt |
|-----------|----------------|---------------------------|--------------|
| (S)       | 55.8           | 55.8                      | 55.1         |
| (D)       | 35.2           | 34.8                      | 34.7         |
| (C)       | 91.0           | 90.6                      | 89.8         |
| (F)       | 2.4            | 2.6                       | 2.7          |
| (M)       | 6.5            | 6.9                       | 7.5          |

| pan-transcriptome | rep | alt | rep+alt |
|-------------------|-----|-----|--------|
| (S)               | 92.2| 11.0| 3.9    |
| (D)               | 6.1 | 85.9| 95.6   |
| (C)               | 98.3| 96.9| 99.4   |
| (F)               | 1.4 | 1.3 | 0.3    |
| (M)               | 0.3 | 1.7 | 0.3    |

**Fig. 4** Sanger sequencing validates the constructed cultivar specific transcriptome. Multiple sequence alignment of NPR1-1 coding sequence obtained from eight *E. coli* colonies (NPR1-1 seq. 1–8) by the Sanger method, assembled short or long-read cv. Rywal transcripts and Phureja DM gene model (Sotub07g016890.1.1). Grey - sequence identity, black - SNPs. The alignment was prepared and visualised with Geneious Prime 2020.1.165.

### Table 6. Assessment of completeness of constructed transcriptomes. Percentage of BUSCOs identified in each transcriptome assembly step. (S): Complete and single-copy BUSCOs %; (D): Complete and duplicated BUSCOs % (C): Complete BUSCOs (S + D) % (F): Fragmented BUSCOs % (M): Missing BUSCOs % rep: representative alt: alternative *

| Database size: 1440. |
Fig. 5 Transcript variants present in pan-transcriptome paralogue gene groups. a) Alignment part of stPanTr_010101 with two PW363-specific SNPs marked by red dots. Such SNPs can be used to design cultivar- or allele-specific qPCR assays. b) Alignment part of stPanTr_074336 showing an alternative splice variant in Désirée, (VdnDe4_33782). Both multiple sequence alignments were made using ClustalOmega v 1.2.1169 and visualized with MView v 1.6660. The remaining alignments can be found in Auxiliary file 21.
It should be noted, that the reconstructed transcriptomes include also the meta-transcriptome stemming from microbial communities present in sampled potato tissues (Auxiliary File 3). We decided not to apply any filter on these transcripts. Inclusion of meta-transcripts makes it possible to also investigate the diversity of plant-associated endo- and epiphytes. The majority of these microbial transcripts will have microbial annotations, facilitating their future removal when necessary for other experiments.

**Cultivar-specific transcriptomes can improve high-throughput sequencing analyses.** Most gene expression studies have been based on either potato UniGenes, assembled from a variety of potato expressed sequence tags (e.g. StGI, POCI), or the reference DM genome transcript models. Studies based on any of these resources have provided useful information on potato gene expression, but also have major drawbacks.

| Reference                  | Désirée | ITAG/PGSC | PGSC | Désirée | ITAG/PGSC | PGSC |
|----------------------------|---------|-----------|------|---------|-----------|------|
| **Mapping statistics/Sample** |         |           |      |         |           |      |
| Number of input reads      | 14,953,659 | 14,610,172 |      |         |           |      |
| Average input read length  | 252     | 252       |      |         |           |      |
| **UNIQUE READS:**          |         |           |      |         |           |      |
| Uniquely mapped reads %    | 74%     | 69%       | 71%  | 73%     | 67%       | 69%  |
| Average mapped length      | 246     | 244       | 246  | 246     | 244       | 246  |
| Number of splices: Total   | 123,490 | 104,658   | 159,842 | 127,414 | 110,423   | 165,595 |
| Number of splices: Non-canonical | 40,960 | 43,764   | 45,750 | 40,898 | 45,459   | 42,978 |
| Mismatch rate per base %   | 0.75%   | 0.94%     | 1.00% | 0.74%   | 0.93%     | 1.00% |
| Deletion rate per base     | 0.05%   | 0.04%     | 0.06% | 0.05%   | 0.04%     | 0.06% |
| Deletion average length    | 3.42    | 3.46      | 2.88  | 3.41    | 3.47      | 2.89  |
| Insertion rate per base    | 0.03%   | 0.02%     | 0.04% | 0.03%   | 0.02%     | 0.04% |
| Insertion average length   | 2.21    | 2.99      | 2.52  | 2.24    | 3.00      | 2.53  |
| **MULTI-MAPPING READS:**   |         |           |      |         |           |      |
| % of reads mapped to multiple loci | 4.9% | 1.95% | 2.74% | 4.86% | 1.68% | 2.52% |
| % of reads mapped to too many loci | 0% | 0% | 0% | 0% | 0% | 0% |
| **UNMAPPED READS:**        |         |           |      |         |           |      |
| % of reads unmapped: too many mismatches | 0% | 0% | 0% | 0% | 0% | 0% |
| % of reads unmapped: too short | 21% | 29% | 27% | 22% | 31% | 29% |
| % of reads unmapped: other | 0% | 0% | 0% | 0% | 0% | 0% |
| **Mapping statistics/Sample** |         |           |      |         |           |      |
| Number of input reads      | 14,755,430 | 44,319,261 |      |         |           |      |
| Average input read length  | 252     | 252       |      |         |           |      |
| **UNIQUE READS:**          |         |           |      |         |           |      |
| Uniquely mapped reads %    | 49%     | 44%       | 46%  | 66%     | 60%       | 62%  |
| Average mapped length      | 245     | 243       | 246  | 246     | 244       | 246  |
| Number of splices: Total   | 95,409  | 77,103    | 115,083 | 346,313 | 292,184   | 440,520 |
| Number of splices: Non-canonical | 33,269 | 31,224   | 31,065 | 115,127 | 120,447   | 119,793 |
| Mismatch rate per base %   | 0.75%   | 0.94%     | 1.01% | 0.75%   | 0.94%     | 1.00% |
| Deletion rate per base     | 0.06%   | 0.05%     | 0.07% | 0.05%   | 0.05%     | 0.06% |
| Deletion average length    | 3.23    | 3.25      | 2.83  | 3.36    | 3.41      | 2.87  |
| Insertion rate per base    | 0.03%   | 0.03%     | 0.04% | 0.03%   | 0.02%     | 0.04% |
| Insertion average length   | 2.27    | 3.15      | 2.56  | 2.24    | 3.04      | 2.54  |
| **MULTI-MAPPING READS:**   |         |           |      |         |           |      |
| % of reads mapped to multiple loci | 3.60% | 1.19% | 1.94% | 4.47% | 1.61% | 2.40% |
| % of reads mapped to too many loci | 0% | 0% | 0% | 0% | 0% | 0% |
| **UNMAPPED READS:**        |         |           |      |         |           |      |
| % of reads unmapped: too many mismatches | 0% | 0% | 0% | 0% | 0% | 0% |
| % of reads unmapped: too short | 47% | 55% | 52% | 30% | 38% | 36% |
| % of reads unmapped: other | 0% | 0% | 0% | 0% | 0% | 0% |

Table 7. Mapping of independent dataset to newly assembled cultivar specific reference transcriptome. Mapping statistics for Désirée leaf samples under drought stress to Désirée, ITAG/PGSC merged and PGSC representative transcriptome sequences is shown. RNA-seq data from Désirée leaf samples under drought stress retrieved from the GEO Series GSE140083 – “Transcriptome profiles of contrasting potato (Solanum tuberosum L.) genotypes under water stress”. No chimeric reads detected. *Relevant % of mapped reads: % of uniquely mapped reads + % of reads mapped to multiple loci.
When using the DM genome as a reference for mapping RNA-Seq reads, the potato research community faces the existence of two overlapping, but not identical, gene model predictions. When using either of available GFFs, we are missing some of the genes known to be encoded in the assembled scaffold. The newly generated merged DM-based GTF helps to circumvent this problem, but even in the merged GTF the cultivar-specific genes and variations are not considered. Differences in expression and important marker transcripts can therefore be missed. In addition, the computational prediction of DM transcript isoforms is incomplete and, in some cases, gene models are incorrectly predicted. On the other hand, the inherent heterogeneity and redundancy of UniGenes or similarly combined transcript sets causes the mapping of short reads to multiple transcripts, thus making the result interpretation more difficult. The cultivar-specific transcriptomes presented here are an improvement, as they include several transcripts missing in the Phureja DM transcriptome. The transcriptomes are also a valuable asset for other high-throughput sequencing applications, such as sRNA-Seq, Degradome-Seq or ATAC-Seq, as we now have more detailed information also on transcript variability within one locus.

The benefit of using cultivar-specific transcriptomes was demonstrated through mapping statistics for Désirée leaf samples under drought stress to Désirée, ITAG/PGSC merged and PGSC representative transcriptome sequences. Taking all three samples together, 5.5% more reads were uniquely mapped to Désirée than the ITAG/PGSC Phureja DM gene models (Table 7). From the reads mapped to Désirée transcriptome, 5.3% mapped to Désirée-specific transcripts and 8.2% to transcripts specific to Group Tuberosum genotypes (Auxiliary Table 826).

Cultivar-specific transcriptomes may also help improve mass-spectroscopy based proteomics. Comprehensive protein databases, obtained from transcriptomic data, offer a higher chance of finding significant targets with peptide spectrum match algorithms, thus enhancing the detection and sensitivity of protein abundance measurements.

Using transcriptomes to inform qPCR amplicon design. Aligning transcript coding sequences from a StPanTr paralogue cluster can be used to inform qPCR primer design in order to study gene expression in different cultivars or specific isoforms by selecting variable transcript regions (Fig. 5). On the other hand, when qPCR assays need to cover multiple cultivars, the nucleotide alignments can be inspected for conservative regions for design.

Limitations. Although the here presented transcriptomes are of higher quality than assemblies produced by a single assembler they are neither comprehensive nor perfectly accurate. Firstly, for cv. Rywal and breeding clone PW363 reference transcriptomes are leaf-specific. Additionally, some transcripts are represented by partial sequence(s) and some might be misassembled. Due to very low expression, some transcripts are missed. One also has to note that some true transcripts with very high similarity to the representative transcript, potentially expressed from distinct alleles differing in the SNP pattern, were removed by filtering. Furthermore, most long non-coding mRNAs are discarded in the EvidenceGene step.

When mapping reads back to the initial transcriptomes a fraction of reads was spliced (Table 3), indicating that the transcripts to which these reads map might originate from splice variants that are not present in the transcriptome. We would however like to point out that contiguity of this reads was not confirmed thus they might represent misassemblies. In Désirée 14,094 transcripts had uniquely mapped spliced reads (7% of the filtered transcriptome), and similarly 7,493 (9%) transcripts in Rywal and 10,220 (6%) in PW363. Low number of non-canonical sites was however confirmed by STARlong mapping: 118 for Désirée, 59 for Rywal and 101 for PW363.

An indication of potential problems with some of the assemblies is that the number of transcripts with length less than 200 nt is 10x higher in constructed cultivar-specific transcriptomes than in DM Phureja PGSC representative transcripts (875, 1377 and 1310 vs. 87 for Désirée, PW363 and Rywal vs. DM Phureja) while the transcript length is about 300 nt shorter on average (Table 5). To check for any misassemblies for genes of interest the users are advised to check the MSA of the paralogue cluster (examples for core transcriptome subset can be found in Auxiliary File 261).

Some of the assembled sequences are also not of potato origin. Vectorscreen implied vector segment contamination in 46 Désirée, 9 PW363 and 31 Rywal sequences, while Centrifuge classified some sequences as originating from common plant pests, among which are the meta-transcriptome sequences of potato microbial communities and the complete genome of the Potato Virus Y (PVY), present due to experimental treatments.

Code availability
All used Bash, Perl, Python, and R/Markdown custom code and scripts complemented with intermediate and processed data (input and output files), and all other supporting information that enable reproduction and re-use are available at FAIRDOMHub under project name _p_stRT (fairdomhub.org/projects/161) under CC BY 4.0 licenses. Data were locally stored in a ISA-tab compliant project directory tree generated using pISA-tree (github.com/NIB-SI/pISA) and uploaded to FAIRDOMHub repository using FAIRDOMHub API and R package pisar (github.com/NIB-SI/pisar).

Received: 27 November 2019; Accepted: 19 June 2020;
Published online: 24 July 2020

References
1. Hardigan, M. A. et al. Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proc. Natl. Acad. Sci. USA 114, E9999–E10008 (2017).
2. Gan, X. et al. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477, 419–423 (2011).
3. Hirsch, C. N. et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26, 121–135 (2014).
4. Jin, M. et al. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Sci. Rep. 6, 18936 (2016).
61. Zagorščak, M. Paralogue clusters selection multiple sequence alignments using ClustalOmega v1.2.1 or MAFFT v7.271 and MView v1.66. FAIRDOMHub https://doi.org/10.15490/FAIRDOMHUB.1.DATAFILE.3691.1 (2020).

62. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729 (2016).

63. Breitwieser, F. P. & Salzberg, S. L. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. Bioinformatics 36, 1303–1304 (2019).

64. Zagorščak, M. Evaluation of constructed reference transcriptomes for presence of the metagenome using Pavian visualisation of the Centrifuge taxonomic classification program output. FAIRDOMHub https://doi.org/10.15490/FAIRDOMHUB.1.DATAFILE.3509.1 (2020).

65. Kearse, M. et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649 (2012).

66. NCBI Sequence Read Archive (SRA), https://identifiers.org/ncbi/insdc.sra:SRP229087 (2019).

67. Zagorščak, M. Read count summary for Désirée drought samples mapped to the representative Phureja DM and Désirée reference transcriptomes. FAIRDOMHub https://doi.org/10.15490/FAIRDOMHUB.1.DATAFILE.3722.2 (2020).

68. Luge, T., Fischer, C. & Sauer, S. Efficient application of de novo RNA assemblers for proteomics informed by transcriptomics. J. Proteome Res. 15, 3938–3943 (2016).

69. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 27, 135–145 (2018).

Acknowledgements
We thank Robin Buell for potato PGSC-ITAG gene model reference table, Thomas Doak and Don Gilbert for advice on using EvidentialGene scripts, Henrik Krnec for BLAST output parser, Andrej Blejec for FAIRDOMhub API usage in R, Špela Baebler for providing DSN Illumina-sequenced samples and Karmen Pogačar for providing Rywal cDNA samples for technical validation. This project was supported by the Slovenian Research Agency (grants P4-0165, J4-4165, J4-7636, J4-8228 and J4-9302), COST actions CA15110 (CHARME) and CA15109 (COSTNET).

Author contributions
M.P., K.G., Ž.R. and M. Zagorščak participated in study design and evaluation of transcriptomes. A.C. provided Rywal illumina-sequenced samples. M.P. collected and pre-processed RNA-Seq datasets and produced CLC, Velvet/Oases and rnaSPAdes de novo assemblies. M. Zouine produced Trinity assemblies. E.T. processed Iso-Seq data. M.P. and M. Zagorščak run tr2aacds scripts and transcriptome annotation, filtering and evaluation software. Ž.R. merged PGSC and ITAG gene models of reference potato genome. M.P., Ž.R. and M. Zagorščak generated the pan-transcriptome. S.T. conducted bacterial transformation and Sanger sequencing analysis. K.G. secured funding, and managed the project. M.P., M. Zagorščak, Ž.R., S.T. and K.G. wrote and edited the manuscript. S.S. helped with interpretation of EvidentialGene results, provided advice on filtering of transcriptomes and language editing. All authors have read and commented the manuscript and approved the final submission. Conceptualization: K.G., M.P. Ž.R., M. Zagorščak; Supervision: K.G.; Project Administration: K.G.; Investigation: M.P., M. Zagorščak; Formal Analysis: M.P., Ž.R., M. Zagorščak; Software: M.P., Ž.R., E.T., M. Zagorščak; Methodology: K.G., M.P., Ž.R., M. Zagorščak; Validation: M.P., Ž.R., S.S., S.T., M. Zagorščak; Data Curation: M.P., M. Zagorščak; Resources: A.C., K.G., M.P., E.T., S.T., M. Zouine; Funding Acquisition: K.G.; Writing - Original Draft Preparation: M.P., M. Zagorščak; Writing - Review & Editing: all authors; Visualization: M.P., M.Zagorščak.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41597-020-00581-4.

Correspondence and requests for materials should be addressed to M.P. or M.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files associated with this article.

© The Author(s) 2020