bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

bioRxiv, 2019, 1–12

doi: xx.xxxx/xxxx
Manuscript in Preparation
PREPRINT

PREPRINT

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

Marko Petek[1,*,†], Maja Zagorščak[1,*,†], Živa Ramšak[1], Sheri Sanders[2], Elizabeth Tseng[3], Mohamed Zouine[4], Anna Coll[1] and Kristina Gruden[1]

[1]Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia and [2]National Center for Genome Analysis and Support (NCGAS), Indiana University, Bloomington, USA and [3]PacBio, Menlo Park, CA, USA and [4]Laboratoire Génomique et Biotechnologie des Fruits, INRA-INP/ENSAT, Castanet-Tolosan, France

[*]marko.petek@nib.si
[*]maja.zagorscak@nib.si
[†]Contributed equally.

Abstract

Background: 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 and its subsequent applications in breeding. **Results:** For the DM genome assembly, two only partially-overlapping gene models exist differing in a unique set of genes and intron/exon structure predictions. First step was to merge and manually curate the merged gene model, creating a union of genes in Phureja scaffold. We next compiled available 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 CLC, Trinity, Velvet, and rnaSPAdes *de novo* assemblers using different settings to test for optimal outcome. In addition, for cultivar Rywal, PacBio Iso-Seq full-length transcriptome sequencing was also performed. Revised EvidentialGene redundancy-reducing pipeline was employed to produce accurate and complete cultivar-specific transcriptomes from assemblers output, as well as to attain the pan-transcriptome. Due to being the most diverse dataset in terms of tissues (stem, seedlings and roots) and experimental conditions, cv. Désirée was the most complete transcriptome (95.8% BUSCO completeness). For cv. Rywal and breeding clone PW363 data were available for leaf samples only and the resulting transcriptomes were less complete than cv. Désirée (89.8% and 89.3% BUSCO completeness, respectively). Cross comparison of these cultivar-specific transcriptomes and merged DM gene model suggests that the core potato transcriptome is comprised of 16,339 genes. The pan-transcriptome contains a total of 95,779 transcripts, of which 54,614 transcripts are not present in the Phureja genome. These represent the variants of the novel genes found in the potato pan-genome. **Conclusions:** Our analysis shows that the available gene model of double-monoploid potato from group Phureja is, to some degree, not complete. The generated transcriptomes and pan-transcriptome represent a valuable resource for potato gene variability exploration, high-throughput -omics analyses, and future breeding programmes.

Key words: *Solanum tuberosum*, *de novo* transcriptome assembly, Iso-Seq PacBio sequencing, short read assembly, full-length sequencing, pan-genome, pan-transcriptome, gene expression, crop plant, EvidentialGene

Compiled on: November 18, 2019.
Draft manuscript prepared by the author.

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Background

At the 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 genes [1]. 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 the gene content, but also to the genetic and epigenetic regulatory elements. Pan-genomes and pan-transcriptomes have been described in the model plant species *Arabidopsis thaliana* [2], and several crop species including maize [3, 4], rice [5], wheat [6] and soybean [7].

While the genome of a double-monoploid clone of *Solanum tuberosum* group Phureja (DM) is available [8], 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 genes [1]. Different breeding programmes have resulted in accumulation of different 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 cultivars [9]. Unfortunately, we could not include these transcriptomes in our pan-transcriptome because the transcriptome assemblies were not publicly accessible.

Based on the DM genome, the PGSC and ITAG annotation consortia [8, 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 on 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 an unified one.

While a combined DM gene set is useful, it is still not as useful as a pan-transcriptome that included 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 [11]. 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 tr2aacds pipeline from EvidentialGene [12] to reduce redundancy across assemblies and obtain cultivar-specific transcriptomes. Finally, we consolidated representative cultivar-specific sequences to generate 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 design of any 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 to assist the breeding programmes.

Data description

Transcriptomic sequences of three potato genotypes, cv. Désirée, cv. Rywal and breeding clone PW363, were obtained from *in-house* RNA-Seq projects and supplemented by publicly available datasets of the same genotypes retrieved from SRA (Table 1).

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 Supplementary Table 1.

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 potato genomics resource [13]. The two models (39,431 PGSC and 34,004 ITAG) were then compared on the basis of their exact chromosomal location and orientation. Genes were considered to be equivalent when the shorter sequence covered at least 70% of the longer sequence. In cases of overlapping gene prediction by both, ITAG IDs were kept as primary. All nontrivial examples of merge (e.g. multiple genes in one prediction model corresponding to one in the other, overlapping of genes in two models, nonmatching directionality of genes and similar) were manually resolved (example in Figure 1). This resulted in a merged DM genome GFF 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 (Supplementary File 1).

Data pre-processing

The complete bioinformatic pipeline is outlined in Figure 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 sequences length allowed was set to 15bp while maximum up to 1kb. Orphaned reads were re-assigned as single-end (SE) reads. Processed reads were pooled into cultivar data sets 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" [14]. To reduce the size of cv. Désirée and cv. Rywal datasets, digital normalization was performed using khmer from bbmap suite v37.68 [15] 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"

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 1. Table of samples used to generate the *de novo* transcriptome assemblies.

Genotype	Sample description[1]	Sequencing platform	Library type[2]	Number of reads[3]	SRA ID
Désirée	PVY inoculated leaves	Illumina	DSN-normalized PE90 unstranded	~54 mio	SRR10070125
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 CCS	SRR8281993, ..., SRR8282008
Rywal	mock and PVY inoculated leaves	Illumina	PE100 strand-specific	~710 mio	SRX6801457, ..., SRX6801468
PW363	PVY inoculated leaves	Illumina	DSN-normalized PE90 unstranded	~104 mio	SRR10070123, SRR10070124
PW363	mock and PVY inoculated leaves	SOLiD	SE50 unstranded	~180 mio	SRR10065430, ..., SRR10065433

[1] PVY, *Potato virus Y*; CPB, Colorado potato beetle
[2] PE, paired-end library (the number stands for read length in nt); SE, 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
[3] For paired-end libraries, pairs are counted as two reads

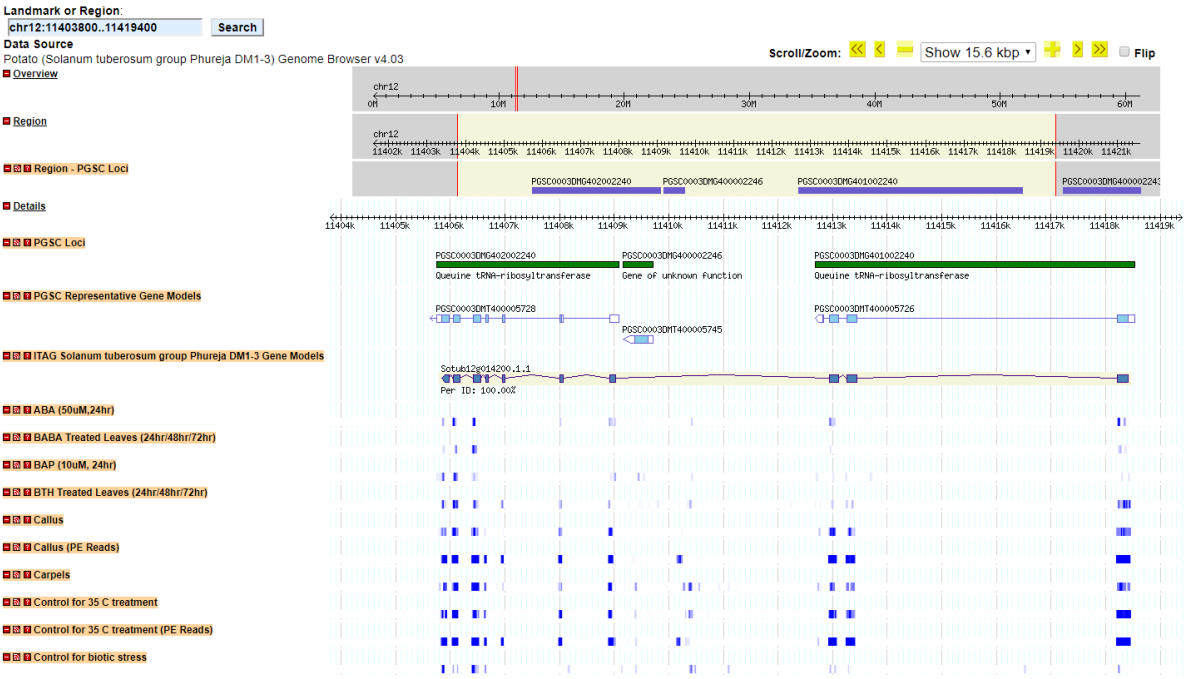

Figure 1. Manual curation example in merged DM genome GFF file generation. Visualisation of region of interest in the Spud DB Genome Browser [13]. ITAG defined Sotub12g014200.1.1 spans three PGSC defined coding sequences (PGSC0003DMT400005728, PGSC0003DMT400005745 and PGSC0003DMT400005726). Corrected transcript was selected based on biological evidence using primary "DM RNASeq Coverage tracks". In cases with missing RNA-Seq data, also other tracks, such as "Other Solanaceae Gene Annotation" and "BLASTP Top Hit", were used. In concrete case, Sotub12g014200.1.1 was preferred due to RNA-Seq evidence in favour of ITAG model.

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

step), clustering isoforms ("cluster" step) and "polishing" step using Arrow consensus algorithm. Only high-quality full-length isoforms were used as input for further steps.

PacBio Cupcake ToFU pipeline

Cupcake ToFU scripts [16] were used to further refine the Iso-Seq transcript set. Redundant isoforms were collapsed with "collapse_isoforms_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.r2013-02-25 [17], Velvet/Oases v. 1.2.10 [18], rnaSPAdes v.3.11.1 [19] 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 much the assembly statistics, therefore we decided to use length 85bp for Illumina datasets of the other two cultivars (Supplementary Figure 1). Parameters k-mer length and bubble size used for Velvet and CLC are given in Table 2. Scaffolding option in CLC and Velvet was disabled. More detailed information per assembly is provided in Supplementary Table 2.

Decreasing redundancy of assemblies and annotation

In order to obtain one clean and non-redundant transcriptome per cultivar, assemblies were first subjected to tr2aacds v2016.07.11 pipeline which grouped and classified transcripts, coding and polypeptide sequences into main (non-redundant), alternative, or discarded (drop) set. Each assembly contributed some proportion of transcripts to cultivar-specific transcriptome sets (Figure 3, Supplementary Figure 1 and Supplementary Figure 2). Both main and alternative sets were merged into initial cultivar reference transcriptomes and used in further external evidence for assembly validation, filtering and annotation steps (Figure 2). *de novo* cultivar-specific transcripts were first mapped to the DM reference genome by STARlong 2.6.1d [20] using parameters optimized for *de novo* transcriptome datasets (all scripts are deposited at FAIRDOMHub project home page). Aligned transcripts were analysed with MatchAnnot to identify transcripts that match the PGSC or ITAG gene models. These transcripts were functionally annotated using the corresponding gene product information in GoMapMan [21]. Domains were assigned to the polypeptide data set using InterProScan software package v5.37-71.0 [22]. For all transcripts and coding sequences, annotations using DIAMOND v0.9.24.125 [23] were generated by querying UniProt retrieved databases (E-value cut-off 10^{-05} and query transcript/cds and target sequence alignment coverage higher or equal to 50%). Assembled initial transcriptomes were also screened for nucleic acid sequences that may be of vector origin (vector segment contamination) using VecScreen plus taxonomy program v.0.16 [24] against NCBI UniVec Database. Potential biological and artificial contamination was identified for cca. 3% of sequences per cultivar. To remove artefacts and contaminants, results from MatchAnnot, InterProScan and DIAMOND were used as biological evidence in a further filtering by *in-house* R scripts (Supplementary scripts 1). Transcripts that did not map to the genome or had no significant hit in either InterPro or UniProt were eliminated from further analysis (Supplementary Table 3, Supplementary Table 4 and Supplementary Table 5). Pajek v5.08 [22], *in-house* scripts, and cdhit-2d from the CD-HIT package v4.6 [25] were used to re-assign post-filtering main and alternative classes and to obtain finalised cultivar-specific transcriptomes (Supplementary scripts 2).

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 (Figure 3, Supplementary Figure 1 and Supplementary Figure 2) to identify optimal parameters for the short-read based assemblers. SOLiD assemblies (Figure 3: 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, for Illumina assemblies, increasing k-mer size in the CLC pipeline produced more complete assemblies according to BUSCO score and more transcripts were selected for the initial transcriptome (Figure 3: 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 (Figure 3: VdnDe1-7). Trinity assembly was comparable in transcriptome contribution and BUSCO score to high k-mer CLC assemblies (Figure 3).

Potato pan-transcriptome construction

Cultivar-specific representative sequences were combined with sequences from the merged DM gene models (non-redundant PGSC and ITAG genes) and subjected to the EvidentialGene traa2cds v2018.06.18 pipeline. Pajek v5.08 [27] and *in-house* scripts (Supplementary scripts 3) were used to retrieve appropriate main and alternative classification of transcripts while also taking discarded sequences in consideration.

138,162 cultivar-specific (57,943 Désirée, 43,883 PW363 and 36,336 Rywal) and 49,322 DM representative sequences were classified into 95,779 main and 91,705 alternative StPanTr transcript sequences. 16,339 main sequences are shared among all three cultivars and the DM clone, while 43,882 sequences are cultivar-specific (i.e. found only in a single cultivar). 17,601 representative sequences from DM clone did not have any match in cv. Désirée, breeding clone PW363 or cv. Rywal (Figure 4, Supplementary Figure 3).

Quality assessment and completeness analysis

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 the transcriptomes via size-based and reference-based metrics, we run TransRate v 1.0.1 [28] on cultivar-specific transcriptomes, prior to and after filtering (Table 4). Comparative metrics for cultivar-specific coding sequences (CDS) were obtained using Conditional Reciprocal Best BLAST (CRBB) [29] against merged 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, and pan-transcriptome), BUSCO v3 [30] scores were calculated using embryophyta_odb9 [26] lineage data (Table 5). At the cultivar-specific transcriptome level, the most diverse dataset in terms of tissues and experimental

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

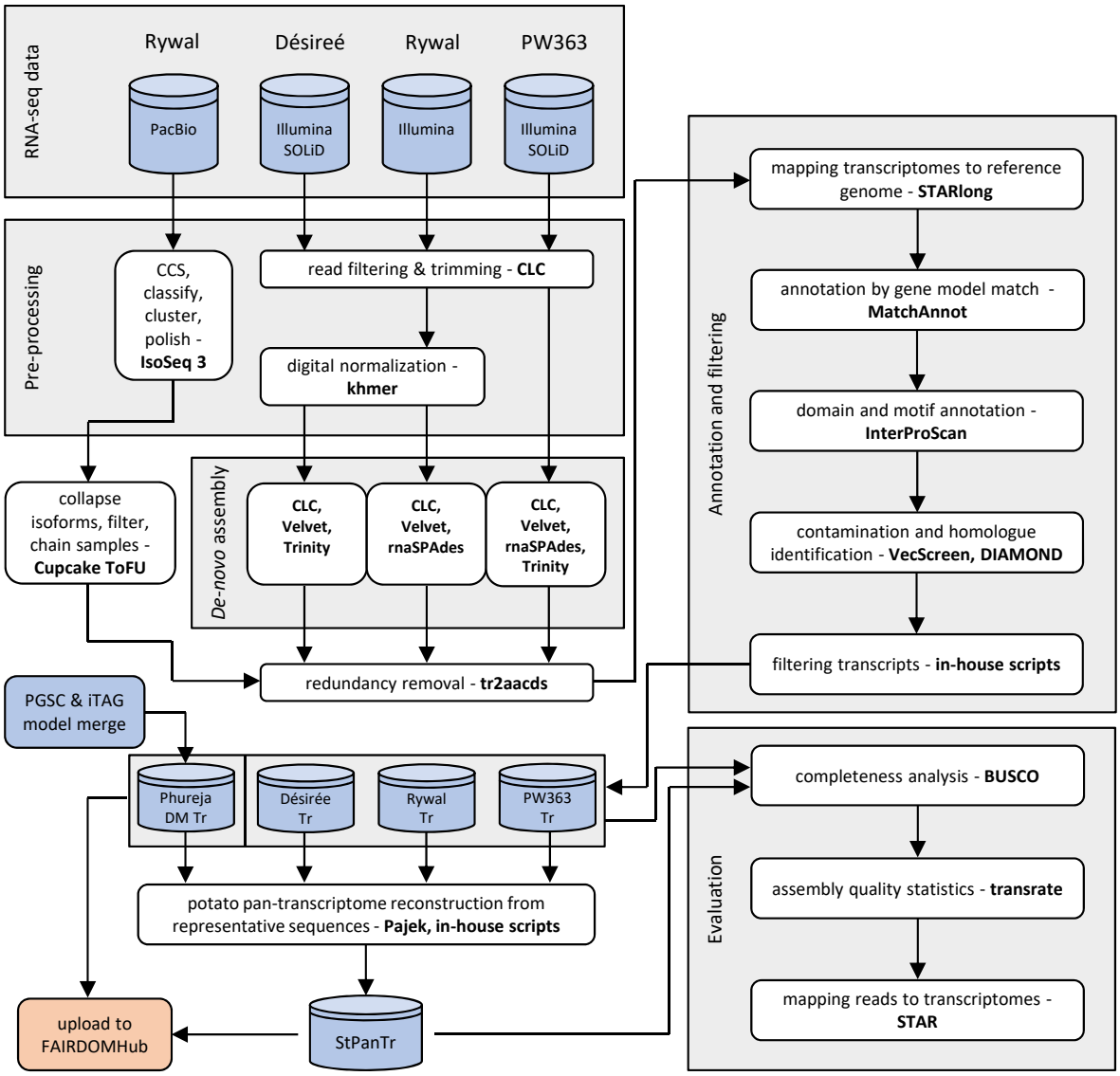

Figure 2. Bioinformatics pipeline for generation of potato transcriptomes. Software used in specific steps are given in bold. Input datasets (sequence reads) and output data (transcriptomes) are depicted as blue cylinders. Data upload steps to public repositories are shaded in orange. Abbreviations: SRA – NCBI Sequence Read Archive, PGSC – Potato Genome Sequencing Consortium, ITAG – international Tomato Annotation Group, CLC – CLC Genomics Workbench, PacBio – Pacific Biosciences Iso-Seq sequencing, Tr – transcriptome, StPanTr – potato pan-transcriptome, tr2aacds – "transcript to amino acid coding sequence" Perl script from EvidentialGene pipeline.

conditions resulted in the highest BUSCO score (cv. Désirée) as expected. Success in classification of main (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). Highest number of fragmented BUSCOs is observed for the breeding clone PW363, what we can probably attribute to the highest number of short-contig assemblies. Furthermore, we can presume how the long-read assembly contributed to the shift in favour of single-copy BUSCOs for cv. Rywal (Table 5) as it has in favour of uniquely mapped reads (Table 3).

To inspect the quality of paralogue cluster assignments, multiple sequence alignments using MAFFT v7.271 [31] were conducted on representative and alternative sequences from paralogue clusters containing 8-16 sequences and containing sequences from each of the four genotypes (Désirée, PW363, Rywal and DM). Alignments were visualized using MView v1.65 [32] (Figure 5, Supplementary File 2). These alignments were our final quality check of success in *de novo* transcriptome assemblies and had helped us also in optimisation of the pipeline. The alignments within groups showed differences that can be attributed to biological diversity, e.g. SNPs and INDELS as well as alternative splicing (Figure 5). We however advise the users to check the available MSA for their gene of interest as some miss-assemblies might still occur (Supplementary File 2, Supplementary scripts 4).

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 2. Parameters used for short reads *de novo* assembly generation.

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* - - transcript discovery as reference	10.0.1	24	50
Désirée	VdnDe8, ..., VdnDe10	SOLiD	Velvet/Oases	1.2.10	23, 33, 43	Default
Désirée	CLCdnDe9, ..., CLCdnDe14	Illumina	CLC *de novo*	9.1	21, 23, 33, 43, 53, 63	85
Désirée	CLCdnDe2, ..., CLCdnDe7	Illumina	CLC *de novo* - - transcript discovery as reference	10.0.1	21, 23, 33, 43, 53, 63	85
Désirée	TDe	Illumina	Trinity	r2013-02-25	25	NA
Désirée	VdnDe1, ..., VdnDe7	Illumina	Velvet/Oases	1.2.10	23, 33, 43, 53, 63, 73, 83	Default
PW363	CLCdnPW1	SOLiD	CLC *de novo*	8.5.4	24	50
PW363	CLCdnPW2	SOLiD	CLC *de novo* - - transcript discovery as reference	9.1	24	50
PW363	VdnPW8, ..., VdnPW10	SOLiD	Velvet/Oases	1.2.10	23, 33, 43	Default
PW363	CLCdnPW3, ..., CLCdnPW44	Illumina	CLC *de novo*	8.5.4	21, 23, 24, 25, 30, 33, 35, 40, 43, 45, 50, 53, 55, 63	50, 65, 85
PW363	CLCdnPW45, ..., CLCdnPW50	Illumina	CLC *de novo* - - transcript discovery as reference	10.0.1	21, 23, 33, 43, 53, 63	85
PW363	SdnPW1	Illumina	rnaSPAdes	3.11.1	43	Default
PW363	TPW	Illumina	Trinity	r2013-02-25	25	NA
PW363	VdnPW1, ..., VdnPW7	Illumina	Velvet/Oases	1.2.10	23, 33, 43, 53, 63, 73, 83	Default
Rywal	PBdnRY1	PacBio Isoseq	Iso-Seq 3, Cupcake ToFU	2017	NAp	NAp
Rywal	CLCdnRY1, ..., CLCdnRY6	Illumina	CLC *de novo*	9.1	21, 23, 33, 43, 53, 63	85
Rywal	CLCdnRY7, ..., CLCdnRY12	Illumina	CLC *de novo* - - transcript discovery as reference	10.1.1	21, 23, 33, 43, 53, 63	85
Rywal	SdnRY1	Illumina	rnaSPAdes	3.11.1	43	Default
Rywal	VdnRY1, ..., VdnRY7	Illumina	Velvet/Oases	1.2.10	23, 33, 43, 53, 63, 73, 83	Default

NAp – not applicable

Re-use potential

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, that are not present in merged Phureja DM gene model. Addition of Iso-Seq dataset in the case of cv. Rywal confirms that long reads contribute to less fragmentation of *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 can be seen in reduction rate for PW363 (24-fold), producing additional short-read assemblies does not contribute as much to the quality of a transcriptome 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 benefitted 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 (Figure 5, Supplementary File 2).

It should be noted, that the reconstructed transcriptomes include also the meta-transcriptome stemming from microbial communities present in sampled potato tissues. 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 se-

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

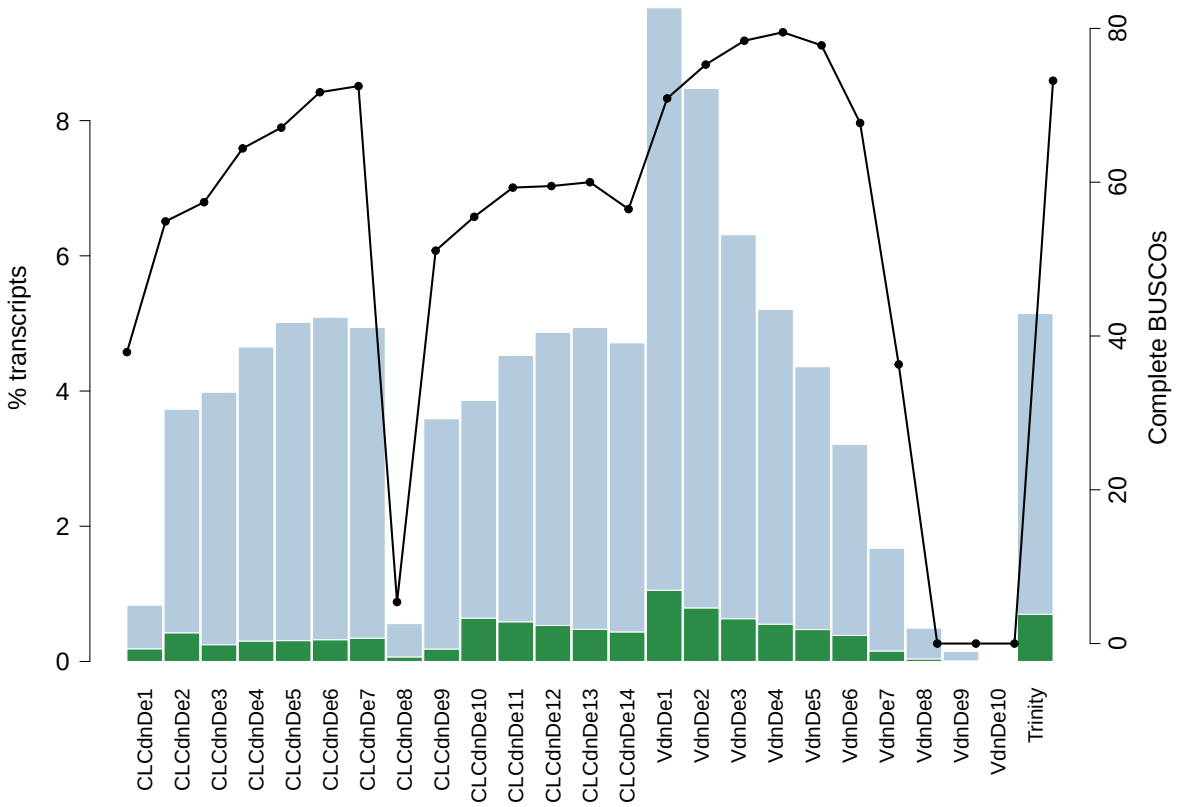

Figure 3. Number of transcripts from *de novo* assemblies contributing to cultivar Désirée, transcriptome and number of complete BUSCOs found in assemblies. Proportion of all contigs in *de novo* assembly (blue bars) and proportion of EvidentialGene okay set (green bars), and the number of complete BUSCOs (dots) using embryophyta_odb9 [26] set are shown. Assembly software abbreviations: CLCdn - CLC Genomics Workbench, Vdn - Velvet. For Rywal and PW363 see Supplementary Figure 2 and Supplementary Figure 1.

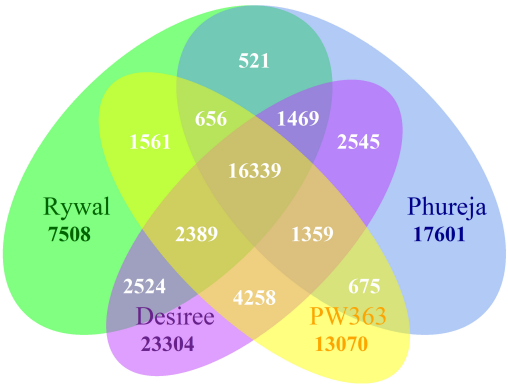

Figure 4. Venn diagram showing the overlap of paralogue clusters in cultivar-specific transcriptomes and merged DM gene model. Only one transcript, i.e. representative, of the StPanTr paralogue cluster is counted. For Phureja, the merged ITAG and PGSC DM gene models were counted.

quence 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 each have major drawbacks.

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 were missing some of the genes known to be encoded in the assembled scaffold. The newly generated merged GFF helps to circumvent this problem. But even when using merged DM-based GFF, 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 similar combined transcript sets causes short reads to map to multiple transcripts and thus makes the interpretation of results more difficult. The cultivar-specific transcriptomes presented here are an improvement as they include some expressed transcripts that are not present in the reference genome and are less redundant than UniGene sets. This is even more so true for different other applications of high-throughput sequencing, such as sRNA-Seq, Degradome-Seq or ATAC-Seq, as we now have more detailed information also on variability of transcripts within one loci which is a requirement for these.

Cultivar-specific transcriptomes may also help improve mass-spectroscopy based proteomics. A more comprehensive database of expressed proteins gives the peptide spectrum match algorithms more chance of obtaining a significant target, thus enhancing the detection and sensitivity of protein abundance measurements [33].

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 expression of specific isoforms or cultivars by selecting

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 3. Assembly accuracy through mapping statistics for initial transcriptomes of individual cultivars.

Mapping statistics/genotype and read type	Désirée SE	Désirée PE	PW363 PE	Rywal SE	Rywal PE
Number of input reads	282,223,241	177,149,132	52,171,015	12,238,901	342,767,035
Average input read length	65	178	179	24	199
UNIQUE READS:					
Uniquely mapped reads number	58,070,281	64,507,790	18,416,487	1,700,383	206,003,021
Uniquely mapped reads %	20.58%	36.41%	35.30%	13.89%	60.10%
Average mapped length	70.12	174.85	175.61	25.56	196.12
Number of splices: Total	220,994	496,170	267,268	2,451	1,700,235
Number of splices: Annotated (sjdb)	0	0	0	0	0
Number of splices: GT/AG	156,753	258,208	105,551	1,486	1,162,885
Number of splices: GC/AG	3,946	10,749	5,693	81	79,495
Number of splices: AT/AC	570	1,486	2,192	1	1,840
Number of splices: Non-canonical	59,725	225,727	153,832	883	456,015
Mismatch rate per base %	0.45%	0.50%	0.53%	2.75%	0.59%
Deletion rate per base	0.03%	0.03%	0.03%	0.00%	0.03%
Deletion average length	2.26	2.72	2.53	1.91	3.02
Insertion rate per base	0.02%	0.02%	0.02%	0.00%	0.03%
Insertion average length	1.46	1.93	1.86	1.49	1.91
MULTI-MAPPING READS:					
Number of reads mapped to multiple loci	193,677,356	98,694,222	29,366,122	2,612,675	108,669,657
% of reads mapped to multiple loci	68.63%	55.71%	56.29%	21.35%	31.70%
Number of reads mapped to too many loci	19,601,924	4,652,918	1,555,704	641,400	1,541,238
% of reads mapped to too many loci	6.95%	2.63%	2.98%	5.24%	0.45%
UNMAPPED READS:					
% of reads unmapped: too many mismatches	0.00%	0.00%	0.00%	0.00%	0.00%
% of reads unmapped: too short	2.62%	5.25%	5.43%	22.94%	7.75%
% of reads unmapped: other	1.24%	0.00%	0.00%	36.57%	0.00%
CHIMERIC READS:					
Number of chimeric reads	0	0	0	0	0
% of chimeric reads	0.00%	0.00%	0.00%	0.00%	0.00%

NAp – not applicable

Table 4. Prior and post-filtering transcriptome summary statistics for potato cultivar-specific coding sequences generated by TransRate.

TransRate metrics	Désirée		PW363		Rywal	
	Pre-filter (initial)	Post-filter	Pre-filter (initial)	Post-filter	Pre-filter (initial)	Post-filter
CONTIG METRICS						
No. sequences	350,271	197,839	273,216	159,278	134,755	79,095
Sequence mean length	504	792	516	775	459	707
No. sequences under 200 nt	125,465	25,330	88,230	17,370	52,653	13,198
No. sequences over 1000 nt	57,679	55,837	44,508	42,571	19,175	18,748
No. sequences over 10000 nt	23	23	3	3	1	1
No. sequences with ORF	132,959	123,546	109,649	100,222	43,870	41,523
'n90	369	444	366	429	351	390
'n50	1,194	1,209	1,110	1,131	1,227	1,218
GC %	41%	42%	42%	42%	42%	42%
Ambiguous nucleotide (N) %	0%	0%	0%	0%	0%	0%
COMPARATIVE METRICS						
No. seq. with CRBB hits*	160,295	138,131	138,443	116,834	66,258	55,239
No. reference seq. with CRBB hits*	29,858	27,642	25,739	23,839	23,549	22,163
coverage50#*	25,991	24,586	21,875	20,620	20,258	19,538
coverage95#*	19,329	18,246	15,664	14,727	14,967	14,470
Reference coverage*	65%	63%	56%	54%	53%	52%

'the largest contig size at which at least 90% or 50% of bases are contained in contigs at least this length
*reference-based summary statistics (merged Phureja DM coding sequences were used as reference)
#proportion of reference proteins with at least N% of their bases covered by a Conditional Reciprocal Best Blast (CRBB) hit

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

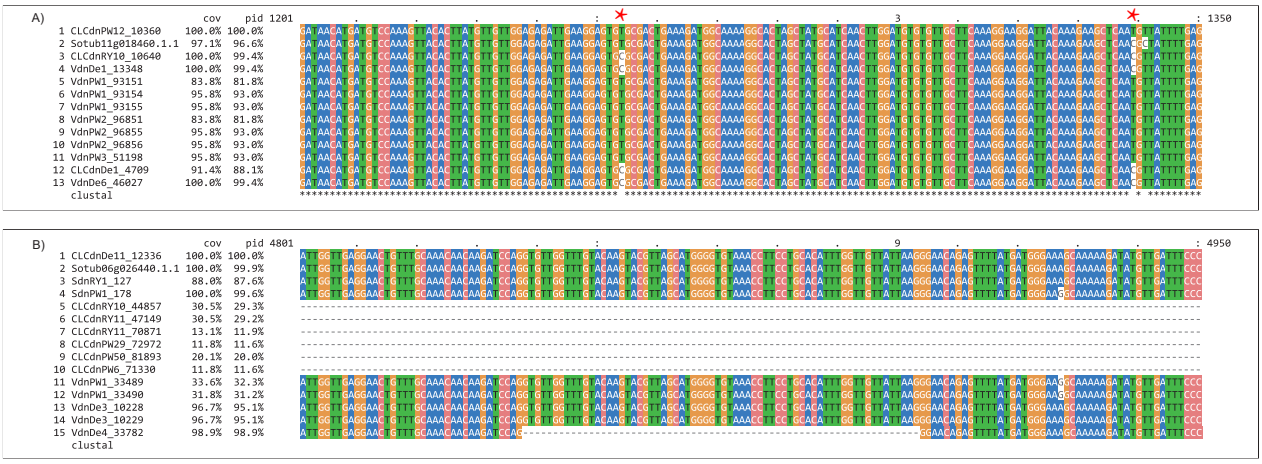

Figure 5. Example alignment of one potato pan-transcriptome paralogue gene group. A) Alignment part of stPanTr_038338 with two PW363-specific SNPs marked by red dots. Such SNPs can be used to design cultivar- or allele-specific qPCR assays. B) Alignment part of stPanTr_007290 showing an alternative splice variant in Désirée, (VdnDe4_33782). Both multiple sequence alignments were made using ClustalOmega v 1.2.1 [34] and visualized with MView v 1.65 [32]. The remaining alignments can be found in Supplementary File 2.

Table 5. Percentage of BUSCOs identified in each transcriptome assembly step.

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
(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
(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
(F)	2.4	2.6	2.7
(M)	6.5	6.9	7.5
pan-transcriptome	**rep**	**alt**	**rep+alt**
(S)	92.4	10.3	3.9
(D)	4.4	87.3	95.6
(F)	2.1	1.1	0.3
(M)	1.1	1.3	0.3

(S): Complete and single-copy BUSCOs %;
(D): Complete and duplicated BUSCOs %
(F): Fragmented BUSCOs %
(M): Missing BUSCOs %
rep: representative
alt: alternative
*Database size: 1440

variable regions of the transcripts (Figure 5). On the other hand, when qPCR assays need to cover multiple cultivars, the nucleotide alignments can be inspected for conservative regions for design.

Conclusion

The transcriptomes present a valuable resource for different applications of high-throughput sequencing analyses as well as for proteomics studies. They will also be a crucial tool to support the breeding programmes of cultivated potato. In future, when new potato RNA-seq datasets become available, the cultivar-specific transcriptomes can be improved and expanded.

Availability of source code and requirements

Lists the following:

- Project ID: _p_stRT
- Project home page: `https://fairdomhub.org/projects/161`;
- Local project data management using "pISA-tree: Standard project directory tree" (ISA-tab compliant) `https://github.com/NIB-SI/pISA`
- pISA-tree – FAIRDOMHub API usage in R: `https://github.com/NIB-SI/pisar`
- Operating systems: Fedora v23, Linux Mint v18.2, Windows 7/8/10
- Programming languages: Bash, Perl, Python, R/Markdown
- License: GPLv3

Availability of supporting data and materials

The GFF file with merged ITAG and PGSC gene models for *S. tuberosum* group Phureja DM genome v4.04 is available at project home page, as the cultivar-specific and pan-transcriptome assembly FASTA and annotation files, custom code described in the manuscript, intermediate and processed data, and all other supporting information that enable reproduction and re-use.

Supplementary tables

1 Supplementary Table S1 – **Detailed sample information table used to generate the *de novo* transcriptome assemblies**. Raw and processed reads summary. Layer _p_stRT/_I_STRT/_S_01_sequences, DOI: 10.15490/fairdomhub.1.datafile.3090.1

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2 Supplementary Table S2 - **Detailed *de novo* assemblies information table**. Primary potato transcriptome assemblies summary listing parameters used for short reads *de novo* assembly generation. Layer _p_stRT/_I_STRT/_S_02_denovo, DOI: 10.15490/fairdomhub.1.datafile.3091.1

3 Supplementary Table S3 - **Désirée biological evidence filtering results**. Output of 1st filtering step by biological evidence for cv. Désirée. Layer _p_stRT/_I_STRT/_S_03_stCuSTr/_A_03.1_filtering, DOI: 10.15490/fairdomhub.1.datafile.3110.1

4 Supplementary Table S4 - **PW363 biological evidence filtering results**. Output of 1st filtering step by biological evidence for breeding clone PW363. Layer _p_stRT/ _I_STRT/_S_03_stCuSTr/_A_03.1_filtering, DOI: 10.15490/fairdomhub.1.datafile.3111.1

5 Supplementary Table S5 - **Rywal biological evidence filtering results**. Output of 1st filtering step by biological evidence for cv. Rywal. Layer _p_stRT/_I_STRT/_S_03_stCuSTr/_A_03.1_filtering, DOI: 10.15490/fairdomhub.1.datafile.3112.1

Supplementary figures

1 Supplementary Figure 1 - **Number of transcripts from *de novo* assemblies contributing to breeding clone PW363, transcriptome and number of complete BUSCOs found in assemblies**. Proportion of all contigs in *de novo* assembly (blue bars) and proportion of EvidentialGene okay set (green bars), and the number of complete BUSCOs (dots) using embryophyta_odb9 set are shown. Assembly software abbreviations: CLCdn - CLC Genomics Workbench, Vdn - Velvet, Sdn - SPAdes. Layer _p_stRT/_I_STRT/_S_03_stCuSTr/_A_02.2_assembly-contribution-count, DOI: 10.15490/fairdomhub.1.datafile.3108.1

2 Supplementary Figure 2 - **Number of transcripts from *de novo* assemblies contributing to cultivar Rywal, transcriptome and number of complete BUSCOs found in assemblies**. Proportion of all contigs in *de novo* assembly (blue bars) and proportion of EvidentialGene okay set (green bars), and the number of complete BUSCOs (dots) using embryophyta_odb9 set are shown. Assembly software abbreviations: CLCdn - CLC Genomics Workbench, Vdn - Velvet, Sdn - SPAdes, PBdn - PacBio. Layer _p_stRT/_I_STRT/_S_03_stCuSTr/_A_02.2_assembly-contribution-count, DOI: 10.15490/fairdomhub.1.datafile.3109.1

3 Supplementary Figure 3 - **Venn diagram showing the overlap of paralogue clusters in cultivar-specific transcriptomes and merged DM gene model**. Representatives and alternatives of the StPanTr paralogue cluster are counted. For Phureja,the merged ITAG and PGSC DM gene models were counted. Layer _p_stRT/_I_STRT/_S_04_stPanTr/_A_04-MSA, DOI: 10.15490/fairdomhub.1.datafile.3096.1

Supplementary files

1 GFF - **merged GFF**. The GFF file with merged ITAG and PGSC gene models for *S. tuberosum* group Phureja DM genome v4.04. Layer _p_stRT/_I_STRT/_S_04_stPanTr/_A_01_evigene_1_3cvs-gffmerged, DOI: 10.15490/fairdomhub.1.datafile.3114.1

2 Supplementary HTML 1 - **Multiple sequence alignments using MAFFT v7.271 and MView v1.65**. Paralogue cluster on representative and alternative sequences, at least one from each of the four genotypes. Clusters containing 8-16 sequences. Layer _p_stRT/_I_STRT/_S_04_stPanTr/_A_04_MSA, DOI: 10.15490/fairdomhub.1.datafile.3116.1

Supplementary scripts

1 in-house scripts 1 - **Evidence filtering**. Corresponding scripts for biological evidence filtering step. Layer _p_stRT/_I_STRT/_S_03_stCuSTr/_A_03.1_filtering, DOI: 10.15490/fairdomhub.1.datafile.3117.1

2 in-house scripts 2 - **stCuSTr paralogue clusters**. Corresponding scripts for StCuSTr post-filtering main (non-redundant) and alternative classes reassignment step. Layer _p_stRT/_I_STRT/_S_03_stCuSTr/_A_03.2_components, DOI: 10.15490/fairdomhub.1.datafile.3118.1

3 in-house scripts 3 - **stPanTr paralogue clusters**. Corresponding scripts for StPanTr main and alternative transcripts classification step. Layer _p_stRT/_I_STRT/_S_04_stPanTr/_A_02_components_1_3cvs-gffmerged, DOI: 10.15490/fairdomhub.1.datafile.3119.1

4 in-house scripts 4 - **MSA**. Corresponding scripts for MSA step. Layer _p_stRT/_I_STRT/_S_04_stPanTr/_A_04_MSA, DOI: 10.15490/fairdomhub.1.datafile.3120.1

Declarations

List of abbreviations

BUSCO: Benchmarking universal single-copy orthologs; CDS: Coding sequence; CLC: CLC Genomics Workbench; CRBB: Conditional Reciprocal Best BLAST; cv.: cultivar; DSN: Duplex-specific nuclease; EST: Expressed sequence tag; Iso-Seq: Isoform sequencing; ITAG: International Tomato Annotation Group; main: representative, non-redundant; NR: Non-redundant; ORF: Open reading frame; PacBio: Pacific Biosciences Iso-Seq sequencing; PE: paired-end; PGSC: Potato Genome Sequencing Consortium; qPCR: Quantitative polymerase chain reaction; RNA-Seq: RNA-Sequencing; SE: single-end; SRA: NCBI Sequence Read Archive; StGI: *Solanum tuberosum* gene indices; StPanTr: *Solanum tuberosum* pan-transcriptome; Tr: transcriptome; tr2aacds: "transcript to amino acid coding sequence" Perl script from EvidentialGene pipeline;

Competing Interests

The authors declare no competing interests.

Funding

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's 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 preprocessed 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. Ž.R. and M.Zagorščak generated the pan-transcriptome. K.G. secured funding, and managed the project. M.P., M.Zagorščak, Ž.R. and K.G. wrote and edited the manuscript. S.S. helped with interpretation of EvidentialGene results, provided advice on

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

filtering of transcriptomes and language editing. All authors have read and commented the manuscript and approved the final submission.

Contributor Role	Authors
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., M. Zagorščak
Data Curation	M.P., M. Zagorščak
Resources	A.C., K.G., M.P., E.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

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 and Špela Baebler for provided DSN Illumina-sequenced samples.

Authors' information

- A.C.: Anna Coll, Anna.Coll@nib.si
- K.G.: Kristina Gruden, Kristina.Gruden@nib.si
- M.P.: Marko Petek, Marko.Petek@nib.si
- Ž.R.: Živa Ramšak, Ziva.Ramsak@nib.si
- S.S.: Sheri Sanders, ss93@iu.edu
- E.T.: Elizabeth Tseng, etseng@pacificbiosciences.com
- M. Zagorščak: Maja Zagorščak, Maja.Zagorscak@nib.si
- M.Zouine: Mohamed Zouine, mohamed.zouine@ensat.fr

References

1. Hardigan MA, Laimbeer FPE, Newton L, Crisovan E, Hamilton JP, Vaillancourt B, et al. Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. Proceedings of the National Academy of Sciences of the United States of America 2017 nov;114(46):E9999–E10008. https://www.pnas.org/content/114/46/E9999.
2. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 2011 sep;477(7365):419–423.
3. Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell 2014 jan;26(1):121–135.
4. Jin M, Liu H, He C, Fu J, Xiao Y, Wang Y, et al. Maize pan-transcriptome provides novel insights into genome complexity and quantitative trait variation. Scientific Reports 2016 jan;6.
5. Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nature Genetics 2018 feb;50(2):278–284.
6. Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee HT, Chan CKK, et al. The pangenome of hexaploid bread wheat. Plant Journal 2017 jun;90(5):1007–1013.
7. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nature Biotechnology 2014 oct;32(10):1045–1052.
8. Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, et al. Genome sequence and analysis of the tuber crop potato. Nature 2011 jul;475(7355):189–195.
9. Liu Y, Lin-Wang K, Deng C, Warran B, Wang L, Yu B, et al. Comparative transcriptome analysis of white and purple potato to identify genes involved in anthocyanin biosynthesis. PLoS ONE 2015 jun;10(6).
10. Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, et al. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 2012;485(7400):635–641.
11. Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience 2019;8(5).
12. Gilbert DG. Genes of the pig, Sus scrofa, reconstructed with EvidentialGene. PeerJ 2019;2019(2).
13. Hirsch CD, Hamilton JP, Childs KL, Cepela J, Crisovan E, Vaillancourt B, et al. Spud DB: A resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. Plant Genome 2014;7(1).
14. Zerbino DR. Using the Velvet de novo assembler for short-read sequencing technologies. Current Protocols in Bioinformatics 2010 sep;(CHAPTER: Unit–11.5.).
15. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research 2015 sep;4:900.
16. Tseng E, cDNA_Cupcake v9.0.1; 2019. https://github.com/Magdoll/cDNA_Cupcake.
17. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 2011 jul;29(7):644–652.
18. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 2012 apr;28(8):1086–1092.
19. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience 2019 sep;8(9).
20. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics 2013 jan;29(1):15–21.
21. Ramšak Ž, Baebler Š, Rotter A, Korbar M, Mozetič I, Usadel B, et al. GoMapMan: Integration, consolidation and visualization of plant gene annotations within the MapMan ontology. Nucleic Acids Research 2014 jan;42(D1).
22. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics 2014 may;30(9):1236–1240.
23. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Methods 2014 jan;12(1):59–60.
24. Schäffer AA, Nawrocki EP, Choi Y, Kitts PA, Karsch-Mizrachi I, McVeigh R. VecScreen_plus_taxonomy: imposing a tax(onomy) increase on vector contamination screening. Bioinformatics 2017 10;34(5):755–759. https://doi.org/10.1093/bioinformatics/btx669.
25. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for

bioRxiv preprint doi: https://doi.org/10.1101/845818; this version posted November 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

clustering the next-generation sequencing data. Bioinformatics 2012 10;28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

26. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Molecular Biology and Evolution 2017 12;35(3):543–548. https://doi.org/10.1093/molbev/msx319.
27. De Nooy W, Mrvar A, Vladimir B. Exploratory social network analysis with Pajek. 3rd edition ed. Cambridge University Press; 2018.
28. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: Reference-free quality assessment of de novo transcriptome assemblies. Genome Research 2016 aug;26(8):1134–1144.
29. Aubry S, Kelly S, Kümpers BMC, Smith-Unna RD, Hibberd JM. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. PLoS Genetics 2014;10(6).
30. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015 oct;31(19):3210–3212.
31. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics 2018;34(14):2490–2492.
32. Brown NP, Leroy C, Sander C. MView: A web-compatible database search or multiple alignment viewer. Bioinformatics 1998;14(4):380–381.
33. Luge T, Fischer C, Sauer S. Efficient Application of de Novo RNA Assemblers for Proteomics Informed by Transcriptomics. Journal of Proteome Research 2016 oct;15(10):3938–3943.
34. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Science 2018 jan;27(1):135–145.