De novo Transcriptome Characterization of Royal Iris (*Iris* section *Oncocyclus*) and Identification of Flower Development Genes

Bar-Lev Yamit[1], Senden Esther[1], Pasmanik-Chor Metsada[2] and Sapir Yuval[1]

[1] The Botanical Garden, School of Plant Sciences and Food Security, G.S. Wise Faculty of Life Science, Tel Aviv University, Israel

[2] Bioinformatics Unit, G.S. Wise Faculty of Life Science, Tel Aviv University, Tel Aviv 69978, Israel

Bar-Lev Yamit abargily@tauex.tau.ac.il, 972-3-6407354, Author for correspondence

Senden Esther esthersenden@gmail.com

Pasmanik-Chor Metsada metsada@post.tau.ac.il

Sapir Yuval sapiry@tauex.tau.ac.il

Acknowledgements

We thank the Technion Genome Center for technical assistance and the Bioinformatics Core Facility at Ben-Gurion University for their assistance in the bioinformatics analysis. We thank N. Kane for the Perl script for mining SSRs from transcriptome sequences. This research was funded by the Israel Science Foundation (grant No. 336/16).

Abstract

Transcriptome sequencing of non-model organisms is valuable resource for the genetic basis of ecological-meaningful traits. The Royal Irises, *Iris* section *Oncocyclus*, are a Middle-East group of species in the course of speciation. The species are characterized with extremely large flowers, a huge range of flower colors and a unique pollination system. The Royal Irises serve as a model for evolutionary processes of speciation and plant ecology. However, there are no transcriptomic and genomic data for molecular characterization.

Here we describe the *de novo* transcriptome sequencing and assembly of *Iris atropurpurea* of the Royal Irises. We employed RNA-seq to analyze the transcriptomes of root, leaf and three stages of flower development. We generated over 195 million paired-end sequencing reads. *De novo* assembly yielded 184,341 transcripts with an average length of 535 bp. At the protein level, a total of 28,709 *Iris* transcripts showed significant similarity with known proteins from UniProt database. Orthologues of key flowering genes and genes related to pigment synthesis were identified, showing differential expression in different tissues. In addition, we identified 1,503 short sequence repeats that can be developed for molecular markers for population genetics in irises.

In the era of large genetic datasets, the *Iris* transcriptome sequencing provides valuable resource for studying adaptation-associated traits in this non-model plant. Although intensive eco-evolutionary studies, this is the first reported transcriptome for the Royal Irises. The data available from this study will facilitate gene discovery, functional genomic studies and development of molecular markers in irises, and will provide genetic tools for their conservation.

Keywords: *De novo*, Assembly, Transcriptome, *Iris*, Flower development

Introduction

Iris is the largest genus in the Iridaceae (Asparagales) with over 300 species (Matthews 1997; Makarevitch et al. 2003). The genus is highly heterogeneous, with species exhibiting a huge range of plant size, flower shape and color and possible habitats (Matthews 1997).

The Royal Irises (*Iris* section *Oncocyclus*) are a Middle-Eastern group of about 32 species and eight intraspecific taxa that are endemics to dry, Mediterranean-type climates and found in the eastern Mediterranean Basin, Caucasica, and central Anatolia (Wilson et al. 2016). Species of section *Oncocyclus* occur in small isolated populations and many are considered rare, threatened or endangered (Shmida and Pollak 2007). These species are characterized by a single large flower on a stem and perennial, short, knobby rhizomes, occasionally with stolons (Sapir and Shmida 2002; Wilson et al. 2016). Plants are diploid with chromosome number of 2n=20 (Avishai and Zohary 1977), a number that is relatively low for *Iris* which range in chromosome number from 2n = 16 in *I. attica* to 2n = 108 in *I. versicolor* (data obtained from Chromosome Count Data Base (Rice et al. 2015)), and in genome size from 2,000 to 30,000 Mbp (Kentner et al. 2003).

The Royal Irises are proposed to be in the course of speciation (Avishai and Zohary 1977; Avishai and Zohary 1980; Arafeh et al. 2002). In recent years, they have emerged as a platform for the study of evolutionary processes of speciation, adaptation, and pollination ecology (Arafeh et al. 2002; Sapir et al. 2005; Sapir et al. 2006; Dorman et al. 2009; Volis et al. 2010; Sapir and Mazzucco 2012; Lavi and Sapir 2015; Wilson et al. 2016; Yardeni et al. 2016). Evolutionary processes and adaptive phenotypes are governed by genetic differences, and thus the study of plant ecology and evolution increasingly depends on molecular approaches, from identifying the genes underlying adaptation and reproductive isolation, to population genetics. In the Royal Irises, however, no genetic and molecular tools are currently available. The primary goal of this work was to generate a reference sequence for the Royal Irises that can be used as a molecular toolbox. While whole genome sequencing of the *Iris* is a challenging task, due to its large genome size, transcriptome sequencing may provide a strong basis for the generation of a genomic resource. Next-generation sequencing (NGS) of RNA (RNA-seq) is a powerful tool for high-throughput gene expression discovery, and for uncovering the genetic basis of biological functions, in non-model organisms (Jain 2011). Only two NGS of transcriptomes were reported in *Iris* (Ballerini et al. 2013; Tian et al. 2015), both representing species from section *Louisiana*, which is a distant clade of *Iris* species (Wilson 2011). Currently, only one NGS-based data is available for the Royal Irises, which is the plastid genome sequence of *Iris gatesii* (Wilson 2014). Previous attempts to transfer molecular tools developed for Louisiana irises to *Oncocyclus* irises, such as development of microsatellite loci or identifying candidate genes, have failed (Y. Sapir, un-published). Nonetheless, low plastid variance among Royal Irises species (Y. Sapir and Y. Bar-Lev, un-published) and lack of nuclear sequences call for a wider set of molecular tools for the Royal Irises. Thus, transcriptome sequencing was chosen as a strategy for developing genetic and genomic tools for the Royal Irises.

The Royal Irises display a remarkable variety of different flower colors on a continuous scale; ranging from extreme dark (i.e., almost black) through purples and pinks to individuals with yellow and white flower (Sapir and Shmida 2002). Flower color plays a significant role in the evolutionary ecology of the Royal Irises (Sapir et al. 2006; Lavi and Sapir 2015). The extent of dark color that exists within the group, and the high variation of color also apparent in some populations, raise questions regarding the genetic control of flower color. In a preliminary study (Y. Sapir, unpublished), we found that this color variation results from variation in the concentration of anthocyanin pigments, namely cyanidin and delphinidin. The anthocyanin biosynthesis pathway (ABP) has remained highly conserved across the angiosperms and the genes in the pathway are functionally characterized in many species (Koes et al. 1994; Zufall and Rausher 2003; Larter et al. 2018). Therefore, the ABP serves as a powerful system for studying gene regulatory processes in plants. The ABP is regulated by a combination of three main transcription factors: R2R3-MYB, basic helix-loop-helix (bHLH) and WD40 class proteins (Hichri et al. 2011). The three transcription factors combine to form the MBW protein complex. When the MBW complex is formed, it binds to the promotor of an ABP structural gene where it regulates gene transcription and ultimately controls anthocyanin production (Mol et al. 1998; Elomaa et al. 2003; Koes et al. 2005; Ramsay and Glover 2005).

In this work, we report the *de novo* assembly of a transcriptome for *Iris atropurpurea* Baker, one of the Royal Irises species. *I. atropurpurea* is a highly endangered plant endemic to Israeli coastal plain (Sapir et al. 2003; Sapir 2016). In recent years this species was well-studied for its morphology (Sapir and Shmida 2002), pollination (Sapir et al. 2005; Sapir et al. 2006; Watts et al. 2013), and speciation and population divergence (Sapir and Mazzucco 2012; Yardeni et al. 2016). All these, make *I. atropurpurea* a good candidate for transcriptome sequencing, that will enable further studies of genetic rescue, population genetics and finding genes underlying phenotypic traits such as flower color.

This is the first reported transcriptome for this section, representing genes expressed in root, young leaf tissue and three stages of flower development. The sequenced *Iris* transcriptome offers a new foundation for genetic studies and enables exploring new research questions.

Methods

Plant material

We used two accessions (genotypes) of *I. atropurpurea*, DR14 and DR8, plants that were brought from a large *I. atropurpurea* population in Dora (32°17'N 34°50'E) in Israel (Figure 1a) and grew at the Tel Aviv University Botanical Garden. Aiming at finding genes related to flower development and to floral traits, we used three different bud developmental stages. We defined bud developmental stage 1 as the earliest detectable bud, where the bud has no color, stage 2 as a bud with beginning of color production, around 1.5 cm in size with the anthers still prominently visible above the petals, and stage 3 as a full colored bud, over 2 cm in size and with the petals covering the anthers (Figure 1b). We collected tissues from the root, young leaf and four buds in three developmental stages (1, two of stage 2 and 3) from DR14, and buds in stages 1 and 2 from DR8. We used buds in stages 1 and 2 from DR8 in order to enlarge the representation of rare or low expressed genes.

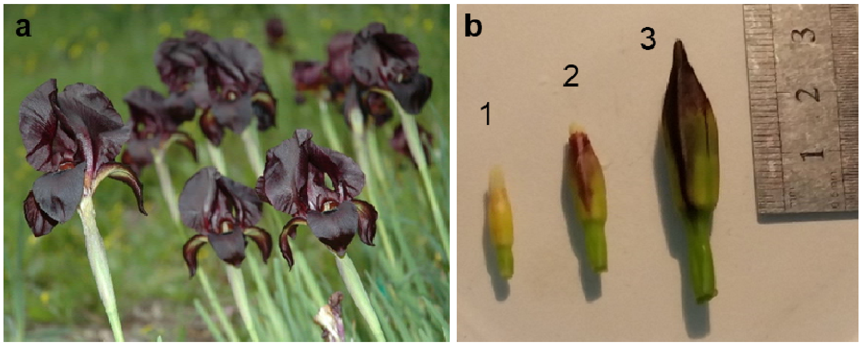

Figure 1. Plant materials used for RNA sequencing. a. *Iris atropurpurea* flower in the field site where collected (Dora). b. Representation of three stages of bud development in *I. atropurpurea*. Bud developmental stages (1 to 3) are defined in text.

RNA isolation and sequencing

We extracted total RNA from all the tissue samples using RNeasy Mini Kit (Qiagen, Hilden, Germany), according to manufacturer's instructions. We measured the quantity and quality of each RNA sample using Qubit fluorometer (Invitrogen) and Bioanalyzer TapeStation 2200 (Agilent Technologies Inc., USA), respectively. Only RNA samples that presented sufficient 260/280 and 260/230 purity and RIN (RNA integrity number) above 8.0 were used for sequencing. RNA was processed by the Technion Genome Center as following: RNA libraries were prepared using TruSeq RNA Library Prep Kit v2 (Illumina), according to manufacturer's instructions, and libraries were sequenced using HiSeq 2500 (Illumina) on one lane of 100 PE run, using HiSeq V4 reagents (Illumina). Sequences generated in this study were deposited in NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under the GEO accession number GSE121786.

De novo assembly and annotation

Raw data were assembled by the bioinformatics core facility (NIBN), Ben-Gurion University, Israel. Quality of the raw sequence reads was estimated using FastQC. De novo assembly of the Iris transcriptome was done using Trinity (version trinityrnaseq_r20140717), with a minimum contig length of 200 base pairs (bp) (Grabherr et al. 2011). We estimated assembly quality using Quast (v. 3.2) (Gurevich et al. 2013). Contigs (isoforms) that are likely to be derived from alternative splice forms or closely-related paralogs were clustered together by Trinity and referred to as "transcripts". The initial reads from each sample were mapped back to the Iris transcriptome that was assembled, using the align_and_estimate_abundance.pl script from trinity pipeline and Bowtie (v. 1.0.0). The number of mapped reads per transcript per sample was counted using RSEM (v. 1.2.25) (Li and Dewey 2011). In order to find the putative genes and function, transcripts were aligned against the UniProt non-redundant protein database (26-09-2016), using BLASTX alignment with an e-value cutoff to < 0.0001 (Altschul et al. 1990).

Differential expression analysis, clustering and functional annotation

To show the gene expression levels in the different organs and bud stages, we used the normalized estimation of gene expression FPKM (fragments per feature kilobase per million reads mapped). FPKM are calculated from the number of reads that mapped to each particular transcript sequence, taking into account the transcript length and the sequencing depth. Transcripts with expression level of less than 2.5 (FPKM) in all tissues were omitted. Expression table was analyzed using Partek Genomic Suite v.6.6 (http://www.partek.com/partek-genomics-suite/) (Downey 2006). Hierarchical cluster analysis was performed to identify expression patterns of transcripts differentially expressed between tissues (as specified in each analysis), by a fold change of at least two. Hierarchical cluster analysis of the full gene list was performed using PROMO (Profiler of Multi-Omics data) (Netanely et al. 2018) with top variance of 2000 genes and 5 clusters, Gene expression was normalized per transcript. PROMO was also used for gene expression visualization of flower development and ABP genes. For the ABP genes heat map the expression was clustered according to expression level per bud stage. Gene Ontology (GO) enrichment analysis of the clusters was obtained using FunRich (Functional Enrichment analysis tool) (Mohashin et al. 2015). Venn diagrams representing number of genes were generated using an online website (http://www.interactivenn.net) (Heberle et al. 2015), after omitting of replicate gene transcripts. The Venn diagram of genes differentially expressed between the three bud stages was generated using Partek Genomic Suite, which accounts for all transcripts and their expression.

The experimental design layout and analysis pipeline are presented in figure 2.

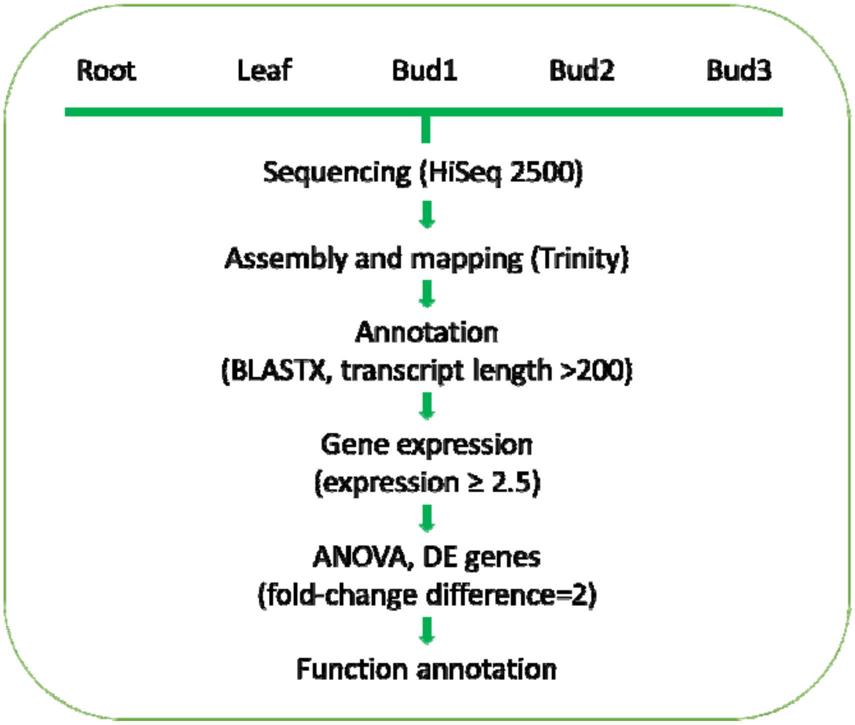

Figure 2. Workflow design of the sequencing, assembly, annotation, gene expression analysis and functional annotation.

SSRs mining

In order to utilize the transcriptome sequenced also for population genetic markers, we searched for short sequence repeats (SSRs; microsatellites) in the assembled contigs. We used a Perl script (find_ssrs.pl; (Barker et al. 2010)) to identify microsatellites in the unigenes. In this study, SSRs were considered to contain motifs with two to six nucleotides in size and a minimum of four contiguous repeat units.

Results

Sequencing and annotation of *Iris* transcriptome

To generate the *Iris* transcriptome, eight cDNA libraries were sequenced: root, leaf and three bud stages from one genet of *I. atropurpurea* (DR14), and buds in stages 1 and 2 from a different genet of the same population (DR8) (Figure 1). We generated a total of 195,412,179 sequence reads. The average GC content of *Iris* contigs was 47% (Table 1 and 2). Reads were of very high quality throughout their length, without evidence of adapter content (Phred score >30).

Using Trinity, we assembled 258,466 contigs (isoforms) longer than 200 bp, which clustered into transcripts, with a total length of 168,049,166 bp. A larger N50 length and average length are considered indicative of better assembly. The longest contig was 27,971 bp and half of the contigs (N50) with more than 500 bp were above 1,312 bp long (Table 1).

To quantify the abundance of contigs assembled, the reads of the separated *Iris* organs were mapped to the assembled contigs, with 125,074,925 mapped reads overall, and an average of 45% reads per tissue that mapped to a unique sequence in the assembled transcriptome.

The length distribution of the assembled contigs revealed that 126,194 (68.46%) contigs ranged from 201 to 500 bp in length; 37,335 (20.25%) contigs ranged from 501 to 1,000 bp in length; 16,282 (8.83%) contigs ranged from 1,001 to 2,000 bp in length; and 4,530 (2.46%) contigs were more than 2,000 bp in length (Figure 3). Descriptive statistics of the sequencing data and transcriptome assembly are summarized in tables 1 and 2.

Table 1. Statistical summary of *Iris* transcriptome sequencing and assembly.

Total reads	195,412,179
Contigs (Isoforms)	258,466
Transcripts	184,341
Transcriptome size	168,049,166
N50 contig size (≥500 bp)	1,312
Largest contig	27,971

Table 2. Descriptive statistics of *Iris* transcriptome samples. GC – Percentage of G or C nucleotides in the sequence.

Plant ID	**Tissue**	**# Paired end sequences**	**#Reads**	**%GC**	**Total mapped reads**	**% Unique mapped reads**
DR14	Root	47,760,556	23,880,278	48	16,671,086	55
	Leaf	52,936,482	26,468,241	47	18,366,659	54
	Bud stage 1	54,806,560	27,403,280	47	16,610,741	40
	Bud stage 2 (a)	51,092,390	25,546,195	47	16,095,948	43
	Bud stage 2 (b)	48,073,466	24,036,733	47	15,363,667	43
	Bud stage 3	59,105,996	29,552,998	47	18,570,119	43
DR8	Bud stage 1	36,949,342	18,474,671	45	10,887,290	40
	Bud stage 2	40,099,566	20,049,783	46	12,509,415	45

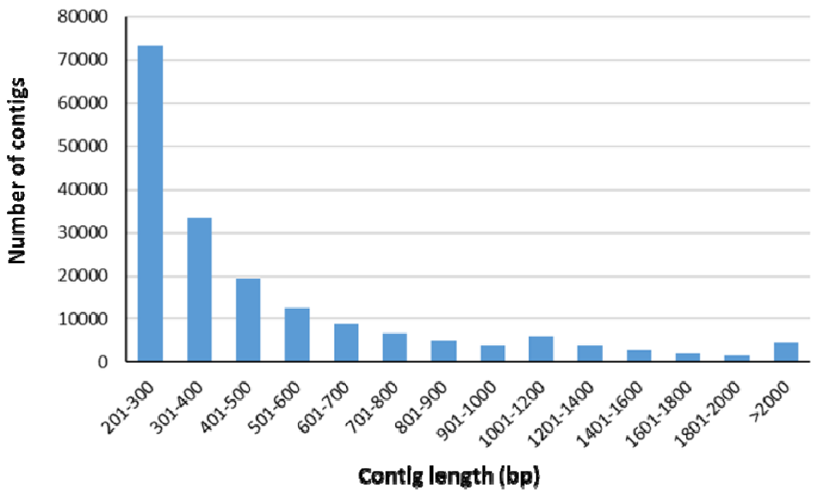

Figure 3. Distribution of contig lengths (in base-pairs) across the assembled contigs from the *Iris* transcriptome.

Using BLASTX search we identified 28,709 transcripts with at least one significant hit. The three most annotated species were *Arabidopsis thaliana* (%), *Oryza sativa* Japonica Group (%) and *Nicotiana tabacum* (%). Interestingly, a considerable number of transcripts were annotated to “non-plant” organisms, most of them to human (*Homo sapiens*, 2. %) (Figure 4).

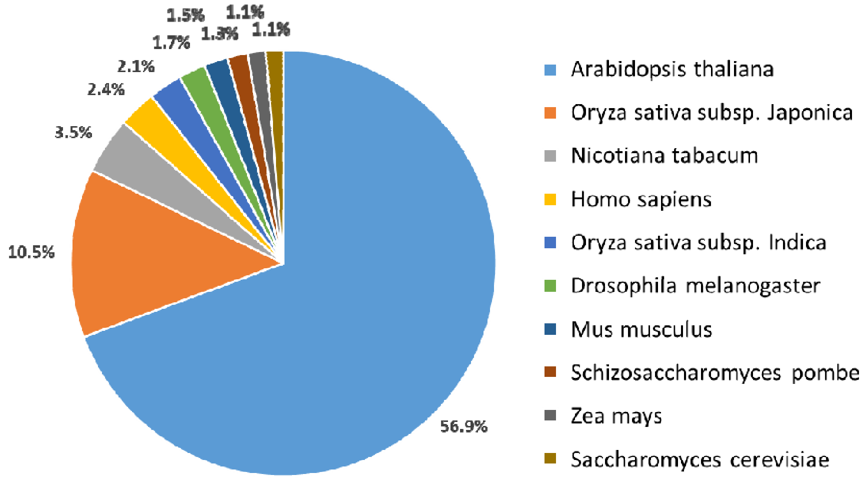

Figure 4. Top 10-hit species distribution of annotated transcripts. Other species represented in the transcriptome had only 1% or less of the transcripts annotated to.

Organ-specific expression analysis

For gene expression analyses, we used only samples from DR14 accession, which had representation of all tissues. We omitted transcripts with expression below 2.5 across samples, leaving 22,804 annotated transcripts (10,935 genes). Organ specific expression of the annotated transcripts revealed 8,025 genes (73%) shared between all tissues. A small portion of the transcripts was organ specific: 337 genes were unique to the root, 135 genes to the leaf, and 948 genes (9%) were expressed only in buds. The number of genes unique to each bud stage decreased gradually during flower development (from 172 in stage 1 to 33 in stage 3) (Figure 5).

Hierarchical cluster analysis of the top 2000 differentially expressed genes revealed 5 clusters with a distinct gene expression pattern for each tissue (Figure S1). The number of differentially expressed genes was highest in stage 1 of bud development (roughly half), and decrease in stages 2 and. Bud 2a and 2b show similar expression pattern.

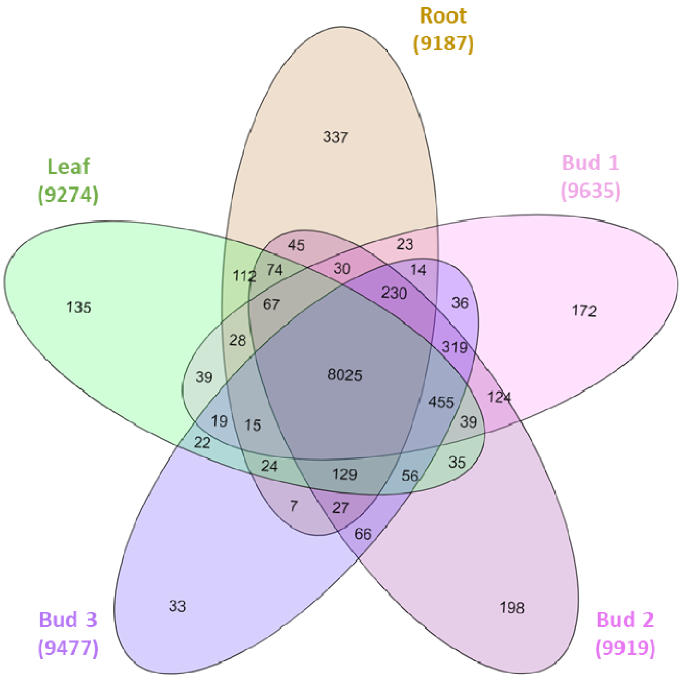

Figure 5. Number of shared and unshared genes (transcripts) between tissues. Bud 2 includes genes expressed in both bud 2a and 2b.

We performed hierarchical cluster analysis on genes differentially expressed between developing buds and vegetative organs (leaves and roots). The analysis revealed three different clusters of transcripts (Figure 6). Cluster #1 (highlighted in pink in figure 6) consisted of 354 transcripts, overexpressed in all buds compared to leaves and root. Gene ontology (GO) analysis of this cluster revealed enrichment of genes related to

RNA modification and transcription regulation (10.1% and 13.2% respectively). A relatively high fraction (4.4%) was of genes related to cell division. Cluster #2 (highlighted in green in figure 6) consisted of 1,757 transcripts, highly expressed in leaves. Employing GO analysis to this cluster revealed gene-expression patterns reflecting stress response such as response to cold stress (4.3%) and response to chitin (2.5%) which is a major component of fungi cell wall (Boller 1995), suggesting existence of endophytic fungi in the leaves. Cluster #3 (highlighted in turquoise in figure 6) contained 1,959 transcripts, with higher expression in the roots. These genes were significantly enriched with genes involved in response to wounding and calcium homeostasis. The highest number of genes (17.2%) was involved in DNA template transcription. The best 25 GO terms for each cluster are presented in figure S2.

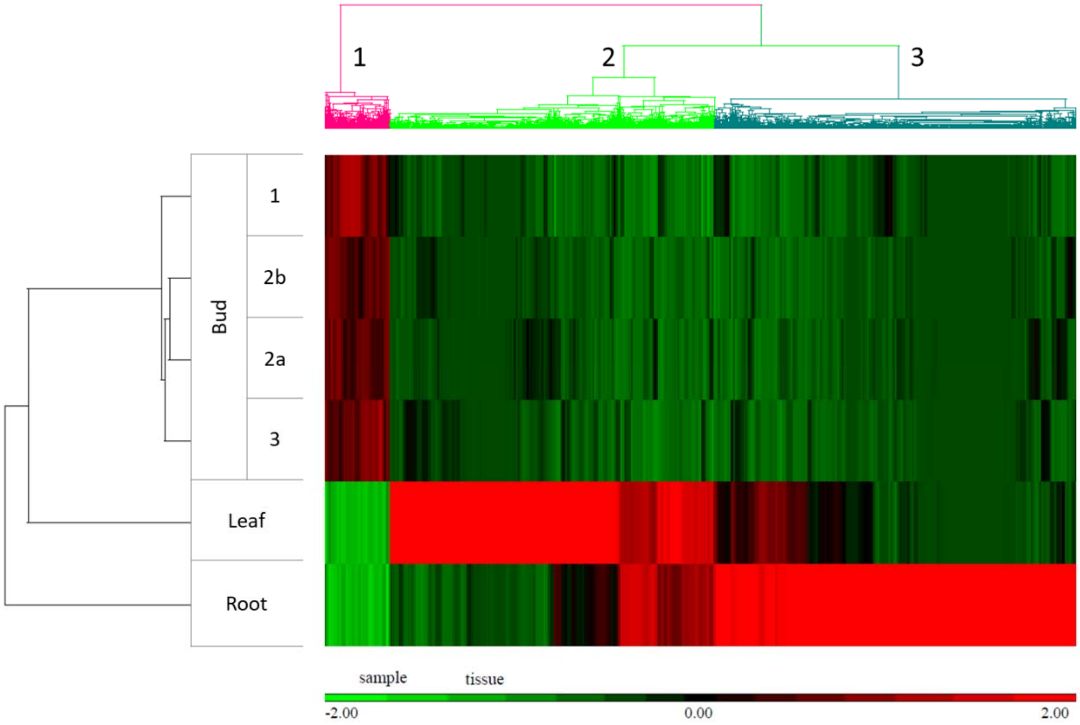

Figure 6. Clustering of genes differentially expressed between the different tissues of *I. atropurpurea* (leaves and roots vs. buds, total of 4,070 genes).

Gene expression in developing buds

Gene expression analysis revealed 10,863 annotated transcripts (6,159 different genes) differentially expressed between bud stages (at least two), 10% of them (1,151 genes) shared between all three lists, i.e. differentially expressed between all bud stages (Figure 7a). Hierarchical clustering of these genes revealed three clusters of expression pattern, cluster 1 which is most distinct and clusters 2 and 3 which have more common features. (Figure 7b).

Cluster #1 (highlighted in orange in figure 7b) consisted of 5,325 transcripts, enriched in stage 1 of the buds. GO analysis of this cluster revealed high enrichment of genes related to transcription (DNA templated) (18.3%) and protein phosphorylation (3.1%). Cluster #2 (highlighted in maroon in figure 7b) consisted of 2,774 transcripts, highly expressed in the two samples of bud stage 2, which were relatively similar. GO analysis in this cluster showed enrichment of genes associated with response to salt stress and cell wall organization. Cluster #3 (highlighted in purple in figure 7b) contained 2,764 transcripts, highly expressed in the bud 3. GO analysis showed that genes in this cluster were enriched with chloroplast organization and protein catabolic processes. Detailed results of GO analysis (25 most enriched) are presented in figure S3.

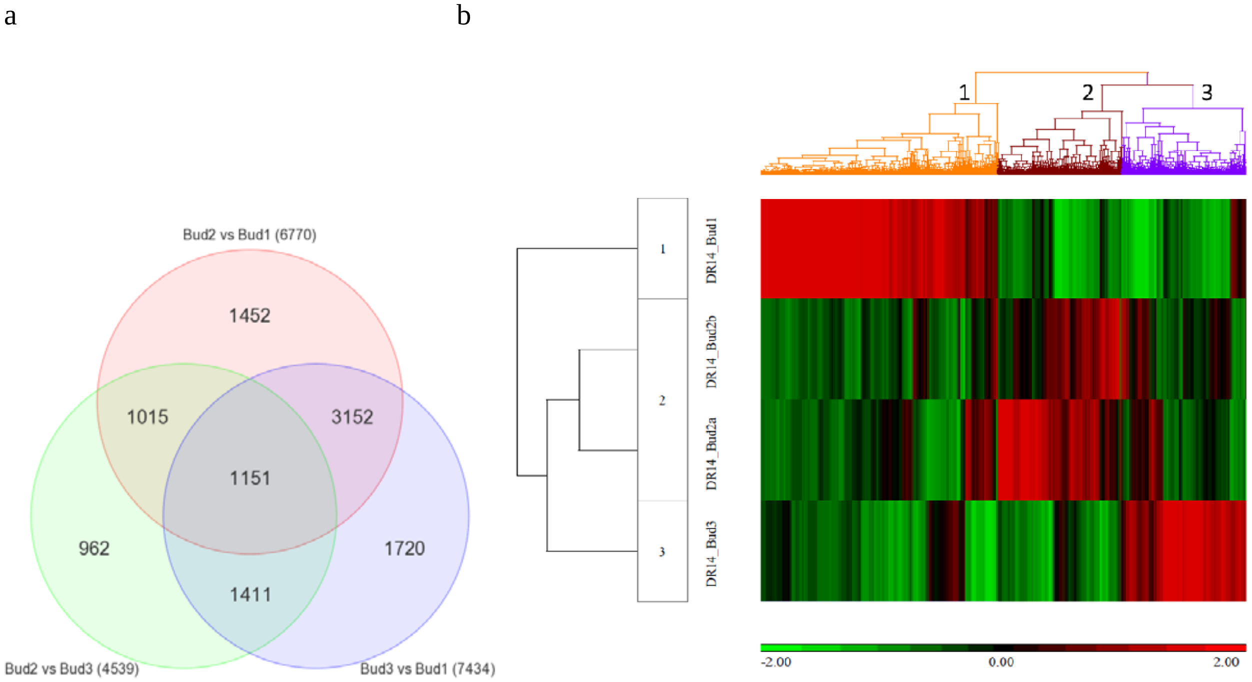

Figure 7. Gene expression in different bud stages. a. Number of genes (transcripts) differentially expressed between bud stages. Three lists of all possible differentially expressed genes were compared. Bud 2 is an averaged expression of bud 2a and 2b. b. Clustering of genes showing differential expression in the three bud developmental stages.

Flower development and anthocyanin biosynthesis genes

In the search for orthologous genes involved in flower development in the irises, we retrieved 721 candidate sequences from the *Arabidopsis thaliana* flower development genes list in the UniProt database (https://www.uniprot.org/). We compared the 6,159 genes differentially expressed between the different buds stages with the *A. thaliana* flower development genes. Only 4.9% of the *Iris* genes (265 genes) were shared with *A. thaliana* genes (Figure 8). The common genes comprise 36% of the UniProt Arabidopsis flower development gene-list.

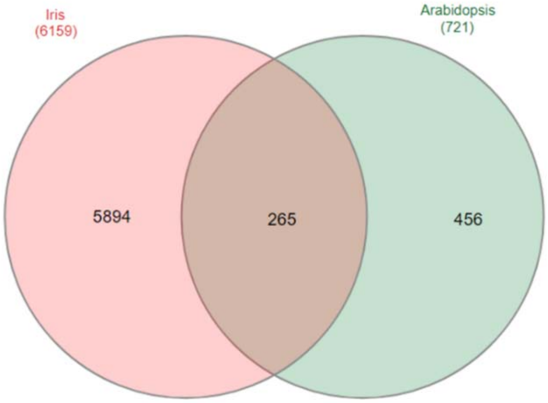

Figure 8. Comparison to *Arabidopsis thaliana* flower development genes. Genes differentially expressed in at least one of the bud stages, in the *Iris* transcriptome, compared to flower development genes in *A.thaliana*, taken from UniProt.

We identified some of the key flowering genes from *A. thaliana* in the *Iris* transcriptome. Five of them belong to the MADS family of transcription factors: MADS2, MADS3 and MADS6, SEPALLATA1 (SEP1) and AGAMOUS (AG) (Bowman et al. 1989), known to have crucial roles in flower organ identity and development (Honma and Goto 2001; Robles and Pelaz 2005). All the orthologues of MADS family were hardly expressed in the vegetative tissues, but their expression patterns varied between bud development stages. SEP1, represented by a single transcript, was the only MADS family gene whose expression did not vary between bud development stages. We identified two genes, APETALA2 (AP2) and WUSCHEL (WUS), which specify the identity of the floral organs (Bowman et al. 1991; Laux et al. 1996). Gene expression analysis of AP2 revealed two transcripts highly expressed in the roots, while all other transcripts present different expression patterns, mostly expressed in the first two stages of bud development (Figure 9). The two transcripts found for WUS, on the other hand, were highly expressed in stage 1 of bud development and not in any other tissue, suggesting it has a role in the first stage of flower organs growth. WUS is involved in the regulation of the AGAMOUS (AG) gene, which induces expansion of floral meristem (Ikeda et al. 2009), potentially ends before stage 2 of the *Iris* bud development.

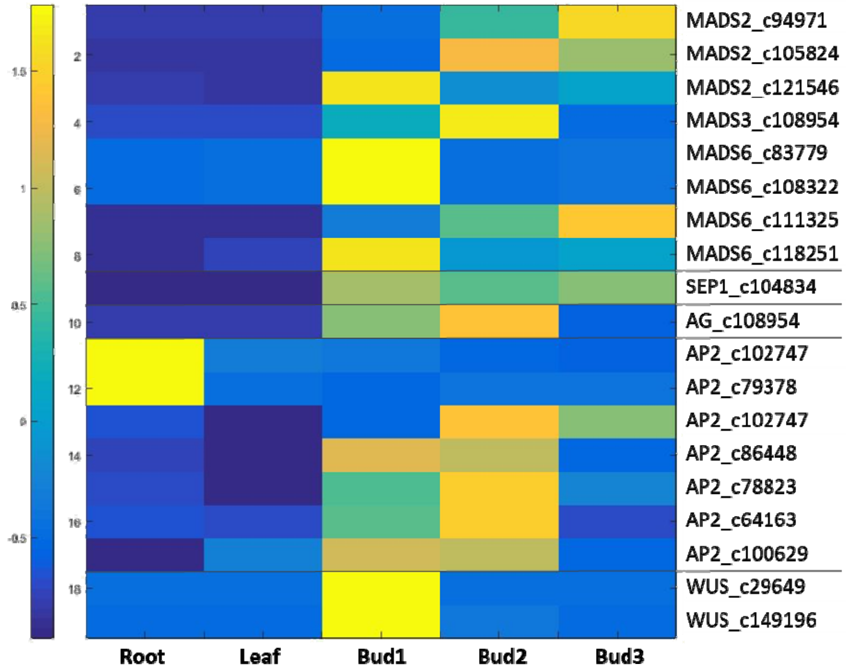

Figure 9. Gene expression of flower development genes in the Iris transcriptome.

Furthermore, we analyzed the transcriptome for the expression of transcripts annotated to structural and regulatory ABP genes. We found 36 transcripts that encode key enzymes (structural genes) in the ABP and 191 transcripts that encode transcription factors (regulatory genes) known to be involved in the regulation of the pathway (Figure 10a). Key enzymes include chalcone synthase (CHS) with 18 transcripts, chalcone isomerase (CHI) with 3 transcripts, leucoanthocyanidin dioxygenase (LDOX) had one transcript, dihydroflavonol 4-reductase (DFR) with 3 transcripts, flavanone 3-hydroxylase (F3H), flavonoid 3' hydroxylase (F3'H) and flavonoid 3'5' hydroxylase (F3'5'H) had 2, 5, and 4 transcripts, respectively. The homology identity levels with known genes ranged between 49.4-90.8%. The regulatory genes bHLH, WD40 and R2R3-MYB transcription factors had 63, 8, and 120 transcripts annotated, respectively. The regulatory genes identity levels ranged between 45.3-90.8%.

Most structural ABP genes were upregulated in bud stages 2 and 3 (Figure 10b). Genes highly expressed in stage 3 were also expressed in stage 2 but to a lower extent, whereas upregulated genes in bud stage 2 had low or no expression in the other stages. The majority of regulatory genes were highly expressed in stage 1, before the initiation of color formation (Figure 10c). In addition to the expression of ABP genes in the buds, we found high expression of some of the genes in the root and the leaf.

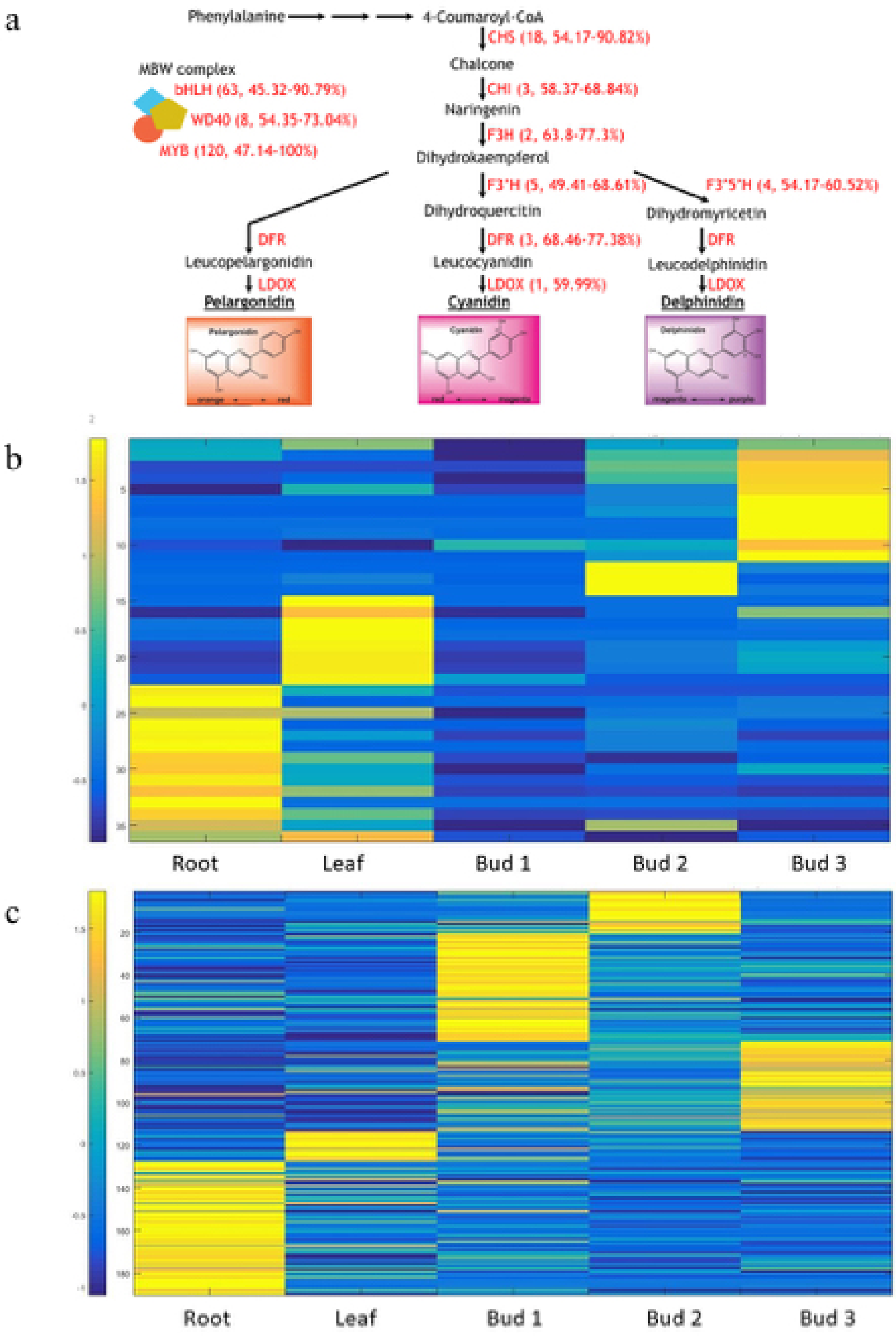

Figure 10. Genes related to anthocyanin pigment biosynthesis pathway. a. A simplified version of the anthocyanin biosynthesis pathway with its structural and regulatory genes. The numbers next to the genes represent the number of transcripts annotated to the gene with the minimum and maximum percentage of identity. b and c are heat maps representing gene expression of ABP genes in the buds and vegetative

tissues of *I. atropurpurea*. The expression levels were normalized per gene and clustered on the basis of expression similarities. b. Expression of transcripts annotated to ABP structural genes (CHI, CHS, F3H, F3'H, F3'5'H, DFR and LDOX). c. Expression of transcripts annotated to ABP regulatory genes (MYB, bHLH and WD40).

Development and characterization of cDNA-derived SSR markers

For the development of new molecular markers, we used all of the 258,466 contigs, generated in this study, to mine potential microsatellites. We defined microsatellites as di- to hexa-nucleotide SSR with a minimum of four repetitions for all motifs. Using a Perl script, we identified 1,503 potential SSRs in 1,241 contigs, of which 263 sequences contained more than one SSR. Only 164 of the contigs containing SSRs had annotation, and were annotated to 115 genes. We assessed the frequency, type and distribution of the potential SSRs (Figure 10). The SSRs included 924 (61.5%) di-nucleotide motifs, 396 (26.4%) tri-nucleotide motifs, 173 (11.5%) tetra-nucleotide motifs and 10 (0.7%) penta-nucleotide motifs. The di-, tri-, tetra- and penta- nucleotide repeats had 8, 30, 37 and 9 types of motifs, respectively. The most abundant di-nucleotide type was GA/TC (254, 16.9%), followed by AG/CT (197, 13.1%) and AT/AT (159, 10.6%). The most abundant tri-nucleotide repeat type was TTC/GAA (37, 2.5%).

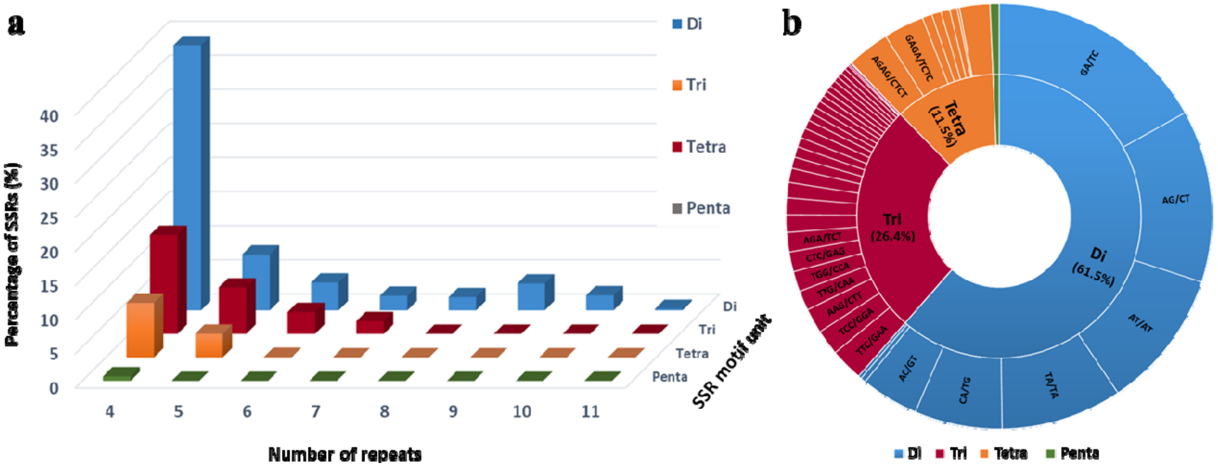

Figure 11. Characterization of SSRs loci found in *Iris* transcriptome. a. Distribution of SSR motif repeat numbers and relative frequency. b. Frequency distribution of SSRs based on motif sequence types.

Discussion

Transcriptome sequencing is one of the most important next generation sequencing tools, and greatly improves our ability to develop genomic resources for non-model organisms. In the past decade, many studies have been using transcriptome *de novo* sequencing and assembly to generate a fundamental source of data for biological research (Meyer et al. 2009; Zhang et al. 2012; Ballerini et al. 2013; Kamenetsky et al. 2015; Tian et al. 2015). In this study, we performed a comprehensive characterization of the transcriptome of *I. atropurpurea*, an important emerging model for understanding evolutionary processes (Arafeh et al. 2002; Sapir et al. 2005; Sapir et al. 2006; Dorman et al. 2009; Volis et al. 2010; Sapir and Mazzucco 2012; Lavi and Sapir 2015; Wilson et al. 2016; Yardeni et al. 2016). We generated 195 million reads that assembled to 184,341 transcripts (contigs, ≥200bp), from leaf, root and three bud developmental stages.

A significant proportion (>65%) of the annotated transcripts were annotated as *Arabidopsis thaliana* or *Oryza sativa*, probably due to higher representation of genomic resources for these species. Surprisingly, out of the transcripts annotated to non-plant organisms, the highest number of transcripts matched to *Homo sapiens*. These transcripts may be attributed to housekeeping genes, which are preserved across all species in eukaryotes, and also due to the highly annotated human genome. The remaining transcripts, which were not matched with a UniProt hit, may be attributed to either not yet identified genes in plants or to species-specific genes in irises.

Most of the genes were expressed in all tissues, probably representing housekeeping genes or genes related to vital functions such as cell growth and proliferation, correlated to the chosen proliferating tissues (young leaf, the tip of the root and developing buds). About 10% of the genes were tissue specific. Large differential organ-specific expression level was found in other plants, such as *Medicago truncatula* (Benedito et al. 2008), *Glycine max* (soybean) (Libault et al. 2010), *Allium sativum* (garlic) (Kamenetsky et al. 2015) and *Arabidopsis thaliana* (Schmid et al. 2005; Aceituno et al. 2008). In our study, the highest number of organ specific genes was in the buds, same as in soybean, garlic, and rice, in which the largest number of tissue-specific genes was also found in the reproductive tissues (Libault et al. 2010; Wang et al. 2010; Kamenetsky et al. 2015). The largest portion of unique genes in buds was in stage 1, which also showed the largest number of over-expressed genes compared to all other tissues and compared to later stages of bud development. This is in concordance with the GO enrichment analysis, showing high enrichment for genes involved in transcription among genes overexpressed in buds compared to leaves and roots, and in genes overexpressed in bud stage 1. Similarly, in garlic, the genes overexpressed in reproductive tissues are enriched with genes involved in DNA replication and regulation of RNA synthesis (Kamenetsky et al. 2015). This suggests high representation of transcription factors (TFs), which play a crucial role in flower development (Smaczniak et al. 2012; Stewart et al. 2016). We identified some of the key TFs regulating flowering, whose expression varied between *Iris* organs, and specifically between bud developmental stages. For example, we identified differential expression of MADS family genes, e.g., MADS-box TFs, SEPALLATA1 (SEP1) and AGAMOUS (AG) (Bowman et al. 1989). The complex action of MADS-domain TFs regulates floral organs identity, and its proteins are essential for flower development (Honma and Goto 2001; Theißen and Saedler 2001; Heijmans et al. 2012). As expected, most of the

MADS-box TFs orthologues were highly expressed in bud stage 1. We also identified the gene APETALA2 (AP2) that specifies sepal identity (Bowman et al. 1989; Robles and Pelaz 2005), and WUSCHEL (WUS) that is required for the identity of shoot and floral meristems (Laux et al. 1996). WUS was suggested to act upstream of other genes (Ikeda et al. 2009), similar to its high expression in stage 1 in *Iris* buds. The expression patterns of these genes in the *Iris* transcriptome suggest that orthologues of flowering genes can be involved in different stages of bud development. Other key flowering genes, which were not identified in our transcriptome, could be genes that were not conserved in irises. Alternatively, these genes may have been expressed in earlier stages of flowering initiation, before the appearance of buds, and thus undetected in the transcriptome. In *Iris lortetii* it was shown that flower organs are mostly expressed in an early stage, about two months prior to stem elongation, when the flower meristem is hidden in the rhizome (Perl 1984). Possibly this is the stage when more flower development genes can be found; however, this stage was not sampled in this study and will be explored in further research.

The characterization of key enzymes in the anthocyanin biosynthesis pathway resulted in 227 annotated transcripts. Focusing on the buds, a large cluster of structural ABP genes was highly expressed in bud stage 3, which is the stage where the color is nearly fully developed. When comparing the expression of the regulatory genes, a large portion was most expressed in stage 1, prior to color production. This may reflect their role as regulators and activators of the ABP, expressed and accumulated prior to the activation and expression of the ABP structural genes. Many ABP genes (both structural and regulatory) were highly expressed in the vegetative tissues. The activity of TF's and structural genes in the root and leaf reflects the production of anthocyanins in vegetative tissues, where they play an important role in plant physiology, development and signaling (Koes et al. 1994; Gould 2004). All three TF's that form the MBW are not exclusively affiliated with the ABP, they are wide spread and are associated with a vast array of different processes, such as regulating cellular diversity within the plant (Ramsay and Glover 2005). Genes associated with pigment synthesis will enable further study on evolution of color variation.

In addition to the characterization of the *Iris* transcriptome, we identified and characterized 1,503 microsatellites (SSR markers), as potential molecular markers. Several studies have already reported the generation of SSR markers from plants transcriptomes (Wang et al. 2010; Garg et al. 2011; Li et al. 2012; Zhang et al. 2012; Mudalkar et al. 2014). Until now, SSRs in irises were reported only for Louisiana and Japanese irises (Tang et al. 2009; Sun et al. 2012); however, these SSRs were not transferable to *Oncocyclus* irises. In our SSRs, the di-nucleotide repeat type was the most abundant motif detected of all repeat lengths. Di-nucleotide SSRs are usually more common in genomic sequences, whereas tri-nucleotide SSRs are more common in RNA sequences (Varshney et al. 2002; Thiel et al. 2003; Luo et al. 2005;

Varshney et al. 2005). Also, tri-nucleotide repeats are more abundant than dinucleotide repeats in plants (Varshney et al. 2005). However, higher number of di-nucleotide repeats in RNA sequences has been reported in Louisiana irises (Tang et al. 2009), and in other plants such as rubber trees (Li et al. 2012) and *Cajanus cajan* (pigeonpea) (Raju et al. 2010). The most abundant di- and tri-nucleotide motifs in *I. atropurpurea* were GA/TC and TTC/GAA, respectively. These results were also coincident with SSRs developed for Louisiana irises, except that the most abundant di- and tri-nucleotide motifs in Louisiana irises were AG/CT and AAG/CTT (Tang et al. 2009).

The *Iris* transcriptome established in this study will increase the molecular resources for irises, which are currently available only for Louisiana and Japanese irises (Tang et al. 2009; Sun et al. 2012), and completely lack for the *Oncocyclus* group. We generated a substantial number of transcript sequences that can be used for discovery of novel genes and genes involved in flower development and pigment production in irises. The numerous SSR markers identified will enable construction of genetic maps and answering important questions in population genetics and conservation. Although *Iris* genetics is still in its early stages, we believe that our transcriptome will significantly support and encourage future evolutionary-genetic research in this ecologically important group.

Compliance with Ethical Standards

Funding: This research was funded by the Israel Science Foundation (grant No. 336/16).

Competing interests: The authors declare that they have no conflict of interests.

Authors' contributions

YBL and YS designed the study and drafted the manuscript. YBL conducted the experimental work. YBL, ES and MPS carried out the bioinformatics analysis. All authors read and approved the final manuscript.

Data availability

All data generated or analyzed during this study are available at NCBI’s Gene Expression Omnibus (GEO), accession number GSE121786.

References

Aceituno FF, Moseyko N, Rhee SY, Gutiérrez RA (2008) The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in *Arabidopsis thaliana*. BMC genomics 9:438

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403-410

Arafeh RM, Sapir Y, Shmida A, Iraki N, Fragman O, Comes HP (2002) Patterns of genetic and phenotypic variation in *Iris haynei* and *I. atrofusca* (*Iris* sect. *Oncocyclus* = the royal irises) along an ecogeographical gradient in Israel and the West Bank. Molecular Ecology 11:39-53

Avishai M, Zohary D (1977) Chromosomes in the *Oncocyclus* Irises. Botanical Gazette 138:502-511

Avishai M, Zohary D (1980) Genetic Affinities among the *Oncocyclus* irises. Botanical Gaztte 141:107-115

Ballerini ES, Mockaitis K, Arnold ML (2013) Transcriptome sequencing and phylogenetic analysis of floral and leaf MIKC(C) MADS-box and R2R3 MYB transcription factors from the monocot *Iris fulva*. Gene 531:337-346

Barker MS, Dlugosch KM, Dinh L, Challa RS, Kane NC, King MG, Rieseberg LH (2010) EvoPipes. net: bioinformatic tools for ecological and evolutionary genomics. Evolutionary Bioinformatics 6:EBO. S5861

Benedito VA, Torres□Jerez I, Murray JD, Andriankaja A, Allen S, Kakar K, Wandrey M, Verdier J, Zuber H, Ott T (2008) A gene expression atlas of the model legume *Medicago truncatula*. The Plant Journal 55:504-513

Boller T (1995) Chemoperception of microbial signals in plant cells. Annual review of plant biology 46:189-214

Bowman JL, Smyth DR, Meyerowitz EM (1989) Genes directing flower development in *Arabidopsis*. Plant Cell 1:37-52

Bowman JL, Smyth DR, Meyerowitz EM (1991) Genetic interactions among floral homeotic genes of *Arabidopsis*. Development 112:1-20

Dorman M, Sapir Y, Volis S (2009) Local adaptation in four *Iris* species tested in a common-garden experiment. Biological Journal of the Linnean Society 98:267-277

Downey T (2006) [13] Analysis of a Multifactor Microarray Study Using Partek Genomics Solution. In: Methods in Enzymology. Academic Press, pp 256-270

Elomaa P, Uimari A, Mehto M, Albert VA, Laitinen RA, Teeri TH (2003) Activation of anthocyanin biosynthesis in Gerbera hybrida (Asteraceae) suggests conserved protein-protein and protein-promoter interactions between the anciently diverged monocots and eudicots. Plant Physiology 133:1831-1842

Garg R, Patel RK, Tyagi AK, Jain M (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Research 18:53-63

Gould KS (2004) Nature's Swiss army knife: the diverse protective roles of anthocyanins in leaves. BioMed Research International 2004:314-320

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075

Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R (2015) InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 16:169

Heijmans K, Morel P, Vandenbussche M (2012) MADS-box Genes and Floral Development: the Dark Side. Journal of Experimental Botany 63:5397-5404

Hichri I, Barrieu F, Bogs J, Kappel C, Delrot S, Lauvergeat V (2011) Recent advances in the transcriptional regulation of the flavonoid biosynthetic pathway. Journal of experimental botany 62:2465-2483

Honma T, Goto K (2001) Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature 409:525

Ikeda M, Mitsuda N, Ohme-Takagi M (2009) *Arabidopsis* WUSCHEL is a bifunctional transcription factor that acts as a repressor in stem cell regulation and as an activator in floral patterning. The Plant Cell 21:3493-3505

Jain M (2011) A next-generation approach to the characterization of a non-model plant transcriptome. Current Science:1435-1439

Kamenetsky R, Faigenboim A, Shemesh Mayer E, Ben Michael T, Gershberg C, Kimhi S, Esquira I, Rohkin Shalom S, Eshel D, Rabinowitch HD, Sherman A (2015) Integrated transcriptome catalogue and organ-specific profiling of gene expression in fertile garlic (*Allium sativum L.*). BMC Genomics 16:12

Kentner EK, Arnold ML, Wessler SR (2003) Characterization of high-copy-number retrotransposons from the large genomes of the louisiana *iris* species and their use as molecular markers. Genetics 164:685-697

Koes R, Verweij W, Quattrocchio F (2005) Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. Trends in plant science 10:236-242

Koes RE, Quattrocchio F, Mol JN (1994) The flavonoid biosynthetic pathway in plants: function and evolution. BioEssays 16:123-132

Larter M, Dunbar-Wallis A, Berardi AE, Smith SD (2018) Convergent evolution at the pathway level: predictable regulatory changes during flower color transitions. Molecular Biology and Evolution:msy117-msy117

Laux T, Mayer KF, Berger J, Jurgens G (1996) The WUSCHEL gene is required for shoot and floral meristem integrity in *Arabidopsis*. Development 122:87-96

Lavi R, Sapir Y (2015) Are pollinators the agents of selection for the extreme large size and dark color in *Oncocyclus* irises? New Phytologist 205:369-377

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323

Li D, Deng Z, Qin B, Liu X, Men Z (2012) De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (Hevea brasiliensis Muell. Arg.). BMC Genomics 13:192

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G (2010) An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. The Plant Journal 63:86-99

Luo M, Dang P, Guo B, He G, Holbrook C, Bausher M, Lee R (2005) Generation of expressed sequence tags (ESTs) for gene discovery and marker development in cultivated peanut. Crop Science 45:346-353

Makarevitch I, Golovnina K, Scherbik S, Blinov A (2003) Phylogenetic Relationships of the Siberian *Iris* Species Inferred from Noncoding Chloroplast DNA Sequences. International Journal of Plant Sciences 164:229-237

Matthews V (1997) A Guide to Species Irises: Their Identification and Cultivation. Edinburgh Journal of Botany 54:367-369

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10:219

Mohashin P, Shivakumar K, Ching-Seng A, Lahiru G, Y.J. QC, A. WN, Dmitri M, M. SO, J. SR, Agus S, Antony B, F. HA, A. SD, T. RM, I. AJ, M. MJ, W. BA, Suresh M (2015) FunRich: An open access standalone functional enrichment and interaction network analysis tool. PROTEOMICS 15:2597-2601

Mol J, Grotewold E, Koes R (1998) How genes paint flowers and seeds. Trends in Plant Science 3:212-217

Mudalkar S, Golla R, Ghatty S, Reddy AR (2014) De novo transcriptome analysis of an imminent biofuel crop, Camelina sativa L. using Illumina GAIIX sequencing platform and identification of SSR markers. Plant Molecular Biology 84:159-171

Netanely D, Stern N, Laufer I, Shamir R (2018) PO-351 Promo: an interactive tool for analysing large multi-omic cancer datasets. ESMO Open 3:A159-A159

Perl A (1984) The control of flowering and the in vitro propagation of *Iris lortetii*. In. The Hebrew University of Jerusalem, Rehovot

Raju NL, Gnanesh BN, Lekha P, Jayashree B, Pande S, Hiremath PJ, Byregowda M, Singh NK, Varshney RK (2010) The first set of EST resource for gene discovery and marker development in pigeonpea (*Cajanus cajanL*.). BMC Plant Biology 10:45

Ramsay NA, Glover BJ (2005) MYB–bHLH–WD40 protein complex and the evolution of cellular diversity. Trends in plant science 10:63-70

Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I (2015) The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. New Phytologist 206:19-26

Robles P, Pelaz S (2005) Flower and fruit development in *Arabidopsis thaliana*. The International Journal of Developmental Biology 49:633-643

Sapir Y (2016) *Iris atropurpurea*. The IUCN Red List of Threatened Species 2016:e.T13161450A18611400

Sapir Y, Mazzucco R (2012) Post-zygotic reproductive isolation among populations of *Iris atropurpurea*: the effect of spatial distance among crosses and the role of inbreeding and outbreeding depression in determining niche width. Evolutionary Ecology Research 14:425-445

Sapir Y, Shmida A (2002) Species concepts and ecogeographical divergence of *Oncocyclus* irises. Israel Journal of Plant Sciences 50:119-127

Sapir Y, Shmida A, Fragman O (2003) Constructing Red Numbers for setting conservation priorities of endangered plant species: Israeli flora as a test case. Journal for Nature Conservation 11:91-107

Sapir Y, Shmida A, Ne'eman G (2005) Pollination of *Oncocyclus* irises (*Iris*: Iridaceae) by night-sheltering male bees. Plant Biology 7:417-424

Sapir Y, Shmida A, Ne'eman G (2006) Morning floral heat as a reward to the pollinators of the *Oncocyclus* irises. Oecologia 147:53-59

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nature Genetics 37:501

Shmida A, Pollak G (2007) Red data book: endangered plants of Israel, Vol. 1. Nature-Parks Jerusalem: Authority Press

Smaczniak C, Immink RGH, Muiño JM, Blanvillain R, Busscher M, Busscher-Lange J, Dinh QD, Liu S, Westphal AH, Boeren S, Parcy F, Xu L, Carles CC, Angenent GC, Kaufmann K (2012) Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. Proceedings of the National Academy of Sciences 109:1560-1565

Stewart D, Graciet E, Wellmer F (2016) Molecular and regulatory mechanisms controlling floral organ development. The FEBS Journal 283:1823-1830

Sun MZ, Li MR, Shi FX, Li L, Liu Y, Li LF, Xiao HX (2012) Genomic and EST□derived microsatellite markers for *Iris laevigata* (Iridaceae) and other congeneric species. American journal of botany 99:e286-e288

Tang S, Okashah RA, Cordonnier-Pratt M-M, Pratt LH, Ed Johnson V, Taylor CA, Arnold ML, Knapp SJ (2009) EST and EST-SSR marker resources for *Iris*. BMC Plant Biology 9:72

Theißen G, Saedler H (2001) Floral quartets. Nature 409:469

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theoretical and applied genetics 106:411-422

Tian S, Gu C, Liu L, Zhu X, Zhao Y, Huang S (2015) Transcriptome Profiling of Louisiana iris Root and Identification of Genes Involved in Lead-Stress Response. International Journal of Molecular Sciences 16:26084

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. TRENDS in Biotechnology 23:48-55

Varshney RK, Thiel T, Stein N, Langridge P, Graner A (2002) In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cellular and Molecular Biology Letters 7:537-546

Volis S, Blecher M, Sapir Y (2010) Application of complex conservation strategy to *Iris atrofusca* of the Northern Negev, Israel. Biodiversity and Conservation 19:3157-3169

Wang L, Xie W, Chen Y, Tang W, Yang J, Ye R, Liu L, Lin Y, Xu C, Xiao J, Zhang Q (2010) A dynamic gene expression atlas covering the entire life cycle of rice. The Plant Journal 61:752-766

Watts S, Sapir Y, Segal B, Dafni A (2013) The endangered *Iris atropurpurea* (Iridaceae) in Israel: honey-bees, night-sheltering male bees and female solitary bees as pollinators. Annals of Botany 111:395-407

Wilson CA (2011) Subgeneric classification in *Iris* re-examined using chloroplast sequence data. Taxon 60:27-35

Wilson CA (2014) The Complete Plastid Genome Sequence of *Iris gatesii* (Section *Oncocyclus*), a Bearded Species from Southeastern Turkey. Aliso: A Journal of Systematic and Evolutionary Botany 32:47-54

Wilson CA, Padiernos J, Sapir Y (2016) The royal irises (*Iris* subg. *Iris* sect. *Oncocyclus*): Plastid and low-copy nuclear data contribute to an understanding of their phylogenetic relationships. Taxon 65:35-46

Yardeni G, Tessler N, Imbert E, Sapir Y (2016) Reproductive isolation between populations of *Iris atropurpurea* is associated with ecological differentiation. Annals of Botany

Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, Zhang Q, Liang X, Li Y (2012) De novo assembly and Characterisation of the Transcriptome during seed development, and generation of genic-SSR markers in Peanut (Arachis hypogaea L.). BMC Genomics 13:90

Zufall RA, Rausher MD (2003) The Genetic Basis of a Flower Color Polymorphism in the Common Morning Glory (Ipomoea purpurea). Journal of Heredity 94:442-448