Comparative transcriptomics of seven *Impatiens* species

Mária Šurinová[1,2]*, Štepán Stočes[3], Tomáš Dostálek[1,2], Andrea Jarošová[2], Zuzana Münzbergová[1,2]

[1]Department of Botany, Faculty of Science, Charles University, Prague, Czech Republic

[2]Institute of Botany, Academy of Sciences of the Czech Republic, Průhonice, Czech Republic

[3]SEQme s.r.o., Dobříš, Czech Republic

*Corresponding author

E-mail: maria.surinova@ibot.cas.cz

Abstract

Impatiens is a genus containing more than 1000 species. Thanks to its size, it is a unique system for studying species diversification in natural populations. This study focused on the characterization of novel transcriptomes from seven *Impatiens* species originating from Nepal. Leave transcriptome of *Impatines balsamina* L.*, I. racemosa* DC.*, I. bicornuta* Wall*, I. falcifer* Hook*, I. devendrae* Pusalkar, *I. scullyi* Hook and *I. scabrida* DC were sequenced and compared. Reads were *de novo* assembled and aligned to 92 035-226 081 contigs. We identified 14 728 orthology groups shared among all the species and 3 020 which were unique to a single species. In single species, 2536-3009 orthology groups were under selection from which 767 were common for all species. Six of the seven investigated species shared 77% of gene families with *I. bicornuta* being the most distinct species. Specific gene families involved in response to different environmental stimuli were closely described. *Impatiens bicornuta* selection profile shared selection on zing finger protein structures and flowering regulation and stress response proteins with the other investigated species. Overall, the study showed substantial similarity in patterns of selections on transcribed genes across the species suggesting similar evolutionary pressures. This suggests that the species group may have evolved via adaptive radiation.

Introduction

Most plant genera contain small number of species with only 57 genera (out of total of more than 13 000) of flowering plants containing more than 500 species (1). Two main mechanisms have been proposed to explain the high diversity of the species rich genera. The first is rapid radiation, which typically occurs in islands or mountains (2). It is described as a process of rapid diversification of an ancestral species via occupation of niches in newly available ecological space (3). The diversification is relatively fast and results in development of new forms adapted to new environments (2). The second proposed process is diversity accumulation in evolution, which in comparison to rapid radiation, is a slower process (4). At the genome level, species differentiation is often linked to polyploidization, genome rearrangement, recombination and complementing mutations (5). These mechanisms are studied for a long time (6–9). On the other hand, studies exploring plant species diversification from the perspective of gene expression are still rare (but see (10). The recent increase in the number of such studies occurred thanks to rise of next-generation sequencing methods allowing to expand whole transcriptome studies to natural, non-model organisms.

From a molecular perspective, species rich genera have been studied very poorly on DNA level and even less on RNA level. Most of the genetic studies comparing multiple species within larger genera are performed to understand species phylogeny by sequencing few genomic regions (11–13). These regions are inadequate to help explaining change in genetic mechanisms leading to adaptation. Knowledge of whole genome or transcriptome sequences is crucial. Gene annotation allows to identify gene functions, regulatory and biochemical pathways for known but also for newly identified genes. Genes are coding information behind the biological networks and this information brings essential insights into protein regulatory

52 interactions that determine biochemical and physiological features of a cell, a tissue and an
53 individual (14).

54 Comparative transcriptomes are mainly available for agricultural species on intraspecies level
55 (15–18). Studies comparing transcriptomes of multiple natural species are rare (19,20). Using
56 RNA sequencing to compare different species has several advantages compared to whole
57 genome sequencing. It allows to study only a regulatory part of the genome allowing to focus
58 on the functional differences among the species and reduce the overall sequencing costs.
59 Comparative studies of natural populations are needed to better elucidate modulation of the
60 genome to understand which mechanisms were involved in species differentiation.

61 Genus *Impatiens* L. (Balsaminaceae) consisting of >1000 species is a good example of a
62 large genus (21). The genus is distributed mainly in tropical regions of the Old World and
63 subtropics (22,23). This genus is widespread from sea level to 4000 m above sea level and
64 several species of the genus have become invasive in different parts of the world (24).
65 Because of its large distribution and high diversity, evolution and adaptation of this genus is
66 widely studied from ecological perspective (25–33). On the other hand, genus *Impatiens* is not
67 deeply studied at the molecular level. No whole-genome sequencing data are available, only
68 six chloroplast genomes have been sequenced so far (34–39) and recently short data report for
69 leaf transcriptome of *Impatience balsamina* was published (40). A combination of broad
70 ecological and genetic research has unique potential to elucidate the importance of
71 environmental factors in species differentiation processes.

72 In this study, we present the first comprehensive analysis of leaf transcriptomes of seven
73 *Impatiens* species. In all seven species, RNA from leaf tissue of select individuals from an
74 altitudinal gradient in Nepal was sequenced using Illumina short reads and assembled *de*
75 *novo* into transcriptomes and annotated. A comparative study followed to characterize the

gene content and identify patterns of selection among orthologous gene families to understand the genetic contribution to the evolution of the genus. Specifically, we asked 1) Do the species differ in transcriptome size and in transcriptome profiles in the de novo sequenced transcriptomes? 2) Do the identified genes under positive selection differ among species?

Material and Methods

Studied species

For our comparative transcriptome study, we used seven *Impatiens* species growing along altitudinal gradient in Nepal, Himalayas: *Impatines balsamina* L.*, I. racemosa* DC.*, I. bicornuta* Wall*, I. falcifer* Hook*, Impatiens aff. I. devendrae* Pusalkar (later referred as *I. devendrae*), *I. scullyi* Hook and *I. scabrida* DC. In recent revision, Akiyama and Ohba (41) suggested that name *I. tricornis* Lindl. should be used instead of *I. scabrida*. In this paper, we are using *I. scabrida* name because of continuity with previous studies. Species of genus *Impatiens* prefer wetter places having low tolerance to long term drought or long exposure to direct sunlight (42). We selected species differing in their altitudinal distribution in order to present species with different temperature niches (31) to investigate possible corresponding changes in transcriptomes. The selection of species was partly limited by the ability of the plants to germinate and survive in all our experimental conditions (see below).

Phylogenetic relationships among the studied species including other members of the genus *Impatiens* is well studied (23,43,44). Published phylogeny (43) defining phylogenetic structure based on sequence and morphological data, identified seven subclades (A-G) within the genus. All species studied here except *I. balsamina* belong to subclade B, *I. balsamina* belongs to subclade G.

Plant material

Seeds of model species were collected from natural populations in autumn 2017 in Nepal (Supplementary Table S1) by M. Rokaya for purpose of our previous studies (26). Permission for the collection of plant material was obtained from the Department of National Parks and Wildlife Conservation in Nepal to Dr. Rokaya. Dr. Rokaya identified collected individuals in collaboration with Dr. Wojciech Adamowski (University of Warsaw). Voucher specimens are available in the Herbarium of the Institute of Botany of ASCR: PRA-20484, PRA-20486, PRA-20488, PRA-20490, PRA-20492, PRA-20494, PRA-20496. Seeds of each species were collected from at least five plants from one population consisting of at least several tens of individuals.

We cultivated the plants in the growth chambers set to the conditions which plants are experiencing in their natural range. Two extremes were selected: (1) cold regime (mean, minimum and maximum temperatures of 12, 6 and 17.5 °C) corresponding to temperatures from March to June at 2,700 m a. s. l., representing the median of the higher altitudinal range of *Impatiens* species in Nepal, and (2) warm regime (mean, minimum and maximum temperatures of 21, 15 and 25 °C) corresponding to temperatures from March to June at 1,800 m a. s. l., representing the median of the lower altitudinal range of *Impatiens* species in Nepal in 2050 as predicted by the global climate model MIRO5C under the greenhouse gas concentration trajectory RCP8.5. The growth chambers were set to 12-h/12-h light/dark cycle, 250 µmol·m-2·s-1 light intensity. Temperature data were obtained from WorldClim database. We used mean temperatures from March to June since it represents premonsoon period when most *Impatiens* species germinate and start to grow. For further details on plant cultivation see (26,31) using the same plants for studies dealing with species trait variation and attractivity to herbivores.

Fourteen samples were collected for RNA extractions (seven investigated species in two regimes). Fully expanded leaves from adult individuals were collected at the stage of flower buds. For each species and temperature regime, the exact age of plants and the time of sampling is different, but all the plants are in the same phenological stage. Leaves were collected during the light period in the same part of the day and immediately frozen in liquid nitrogen.

RNA sequencing

RNA was extracted according to (45) with the following changes. 2.0 M sodium chloride and 25 mM ethylenediaminetetracetic acid were adjusted to 1.4 M sodium chloride and to 20 mM ethylenediaminetetracetic acid in extraction buffer. Despite these adjustments and multiple washing steps, the extracted RNA solution still contained contaminants with ratio 260/230 nm lower than 1.5 and thus had to be purified with glycogen (RNA grade, Thermo Fisher Scientific). Purification with glycogen was performed according to manufactures protocol with changes. Incubation step was performed in room temperature not at -20°C. Thanks to this adjustment, the extracted RNA precipitated but the contaminations remained desolved in the working solution. After precipitation, washing step was performed at least 2×. RNA extraction was performed in technical duplicates, so 28 RNA extractions was used for next step. Library preparation with polyA selection was performed with Dynabeads® mRNA Purification Kit and NEBNext Ultra Directional RNA Library Prep Kit for Illumina according to manufacture protocol. Indexed cDNA libraries were sequenced in one run following manufacturer's instructions to generate paired-end, 150-base reads for reference transcriptomes and in two interdependent runs for single-end, 100bp (base pair) reads. Sequencing runs were performed on a HiSeq 4000 (Illumina). The datasets generated and analyzed during the current study are available in the NCBI database repository TSA: GJBZ00000000, GJBY00000000, GJBR00000000, GJBQ00000000, GJBP00000000, GJBO00000000, GJBN00000000.

Transcriptomes assembly

Transcriptome assembly was performed based on recommendations for de novo assembly of non-model species (46). The quality of the reads and the effects of trimming were assessed using FastQC v0.11.8 (47). Raw reads were trimmed using Trimmomatic v0.39 (48) to remove low quality bases (phred score > 30), adapter sequences, and other sequencing artifacts. After this filtering step, biological replicates were merged to one for each individual, reads were then combined between libraries of the same species. In next step, we filtered out reads based on overall score (phred score > 25) per read, other parameters run on default settings. The erroneous k-mers from trimmed and filtered Illumina paired-end reads were corrected using Rcorrector v1.0.4. (49).

The corrected paired reads were assembled using Trinity v2.8.4 (50) (minimal contig_length: 300; group_pairs distance: 250; minimal kmer_cov: 2) separately for each species (2 samples together). First, the overlapping reads are combined into contigs with a minimum length of 300bp and at least 2 reads to be assembled. Subsequently, these contigs were reassembled into longer sequences of so-called unigenes. Because multiple samples from the same species were sequenced, the longest sequences were grouped into clusters based on the detection of splice variants and abundance. But because splice variants may be incorrectly classified as paralogs (as they can be assembled into different contigs but with the same component name in the assemblies) we retained only the longest isoform for each transcript, thus reducing this assembly error and making sure that each gene was represented by only one transcript. Evaluation of the structure of the generated assemblies was done with the QUAST software (51).

Functional annotation and gene ontology analysis

Assembled non-redundant unigenes were functionally assessed for transcriptome completeness using Benchmarking Universal Single-Copy Orthologs (BUSCO (52)), which uses the entire embryophyte dataset, which represents the evolutionary informed near-universal single copy orthologs from OrthoDB v9. For quality assessment, we ran RSEM-EVAL from DETONATE software package (53). Based on RSEM-EVAL *contig impact score*, we removed all contigs with negative score from the assemblies and reran in RSEM-EVAL. Annotation was performed using the Trinotate pipeline (https://trinotate.github.io/). The coding regions in the transcripts and their most likely candidates for the peptides with the longest open reading frames (ORFs) were predicted using Transdecoder v.2.0.1 (54). Only the putative ORFs that were at least 100 amino acids in length were retained. Subsequently, sequence homology between predicted protein sequences and the NCBI non-redundant protein database (Nr) (ftp://ftp.ncbi.nih.gov/blast/db/) and Swissprot-Uniprot database was searched. The functional annotation was achieved using Hmmer v.3.1b1, a protein domain identification (PFAM) software (55) and Rnammer v.1.2 to predict ribosomal RNA (56). Sequences with a match in either NCBI non-redundant protein database (Nr) or Swissprot-Uniprot database were annotated with gene ontology (GO) terms using the PANTHER (protein annotation through evolutionary relationship) classification tool (57).

Gene families, selection and candidate genes

We used the program OrthoFinder v1.1.2 (58) to identify orthogroups and orthologs of proteins among our 7 assemblies using Basic Local Alignment Search Tool (BLAST) all-v-all (self and reciprocal BLASTs simultaneously) and to derive rooted gene trees for all orthogroups and to identify all of the gene duplication. OrthoFinder analysis was conducted for all pair-wise comparisons, for all 7 species assemblies in order to identify putative orthologs within the

current datasets. We used the outputs from OrthoFinder to determine the number of overlapping (shared across species) transcripts across the 7 assemblies.

The protein sequences associated with the ortholog group members were aligned using MAFFT v7.407 (59) and the corresponding coding sequence was matched and retrieved to ortholog alignment with PAL2NAL v14.1 (60) . Subsequently the alignment was used to construct an maximum likelihood phylogenetic tree using RAxML (Randomized Axelerated Maximum Likelihood) v8.2.12 (61). The algorithm selected for this study conducted a rapid Bootstrap analysis and searched for the best-scoring maximum likelihood tree in one single program run. The selected substitution model was generalized time reversible (GTR) GAMMA. To maximum likelihood to statistically estimate dN and dS was used the CODEML program within the Phylogenetic Analysis by Maximum Likelihood (PAML) package (62). Which uses the observed changes present in the codon alignments from PAL2NAL, given the phylogenetic tree constructed by RAxML. CODEML calculates the likelihood of the observed changes resulting from two models of evolution, only one of which allows for the possibility of detecting positive selection (dN/dS > 1). GO terms were assigned for positively selected orthogroups as described above.

Results

De novo transcriptomes and assembly

RNA-seq libraries representing leaf transcriptomes yielded between 7 981 717 and 15 552 519 paired end reads per individual, for 2 individuals per species cultivated in contrasting temperature regimes. Between 1 and 26% of raw reads were removed due to quality filtering (quality score < 30). For each of the seven species, high quality RNA-seq datasets, which

contained between 53 742 722 and 78 519 184 pair end reads were used for the assembly (Table 1).

Table 1. Summary statistics of sequencing data and de novo transcriptome assembly of seven *Impatiens* species.

	I. balsamina	*I. racemosa*	*I. bicornuta*	*I. scullyi*	*I. scabrida*	*I. falcifer*	*I. devendrae*
total number of reads	64 487 224	63 123 234	68 824 986	78 519 184	59 245 590	70 058 778	64 652 184
total length of reads (bp)	9 157 185 808	8 837 252 760	9 635 498 040	10 992 685 760	7 968 531 855	9 703 140 753	8 835 836 304
% from raw reads	76	85	90	90	80	78	79
total number of contigs	190 736	194 470	226 081	136 472	94 707	161 825	128 266
total length of assembly (bp)	178 316 706	166 381 257	196 656 333	135 964 285	95 826 671	152 280 780	124 053 726
mean length of assembly (bp)	935	856	870	996	1012	941	976
largest contig	221 761	199 594	64 849	245 534	69 423	27 167	48 295
N50	1 351	1 284	1 336	1 599	1 452	1 469	1 461
L50	31 753	31 996	39 334	22 613	20 285	30 184	25 235
GC (%)	38.99	38.64	38.02	40.80	40.50	38.35	39.10

Abbreviations: N50 - the sequence length of the shortest contig at 50% of the total genome length. L50 - the smallest number of contigs whose length sum makes up half of genome size. GC%- is the percentage of nitrogenous bases in a RNA molecule.

The de novo assemblies (using Trinity) ranged between 92 035 and 226 081 contigs for the seven species (Table 2). More than 80% of the input reads mapped back to their assembled transcriptome, which is an indication of a good quality assembly. Similarly also RSEM-EVAL results (Table 2) showed proper assembly.

Table 2. Detonate RSEM-EVAL evaluation of results for each species

Species	Number of contigs	*RSEM-EVAL* score	Number of contigs with no read aligned to	Number of alignable reads
I. balsamina	190 736	-5 610 093 097.24	15 036	16 115 961
I. racemosa	194 470	-6 480 726 545.38	20 400	16 980 323
I. bicornuta	226 081	-8 102 571 944.57	25 347	17 092 225
I. scullyi	136 472	-8 759 635 212.49	12 326	19 586 591
I. scabrida	94 707	-7 102 916 062.01	11 122	12 472 536
I. falcifer	161 825	-7 492 222 641.16	18 948	18 476 547
I. devendrae	141 792	-7 866 313 302.92	18 852	12 712 480

Functional annotation, gene ontology analysis and quality assessment

To perform functional annotation, the filtered assembly was submitted to the Trinotate pipeline and we predicted open reading frames (ORFs) with Transdecoder. Between 42.9 and 68.7% transcripts of *Impatiens* species were identified, resulting into 52.3- 72.4% predicted protein coding genes (Fig 1). A protein gene set completeness assessment (BUSCO pipeline) showed that majority of *Impatiens* core genes had been successfully recovered in assemblies (Table 3). Only between 13.6 and 22 % of the *Impatiens* single-copy orthologs were classified as missing, suggesting good coverage and high quality of the assembly of the protein-coding transcriptomes.

Fig.1. Annotation summary. Quantity of assembled transcripts with identified ORF by Transdecoder (Trinotate pipeline). Percentage values are showing proportion of annotated and undetermined transcripts for each species.

Table 1. BUSCO- an analysis of assembly completeness.

	I. balsamina	*I. racemose*	*I. bicornuta*	*I. scullyi*	*I. scabrida*	*I. falcifer*	*I. devendrae*
Complete genes (%)	73.5	76.7	79.1	77.0	64.7	80.0	72.4
Single-copy genes (%)	40.5	40.9	31.0	42.5	31.9	34.9	33.4
Duplicated genes (%)	33.0	35.8	48.1	34.5	32.8	45.1	39.0
Fragmented genes (%)	10.3	7.2	6.2	8.1	13.3	6.0	9.7
Missing genes (%)	16.2	16.1	14.7	14.9	22.0	14.0	18.0

Annotated genes were assigned to GO terms, with the highest proportions of mapped GO terms for the current *Impatiens* transcriptomes related to binding (~28.3%) and catalytic activity (~50.4%) under "Molecular Function", cellular (~32%) and metabolic (24%) processes under "Biological Process", and cell part (~48.5%) and organelle (~36.5%) under "Cellular Component" (Fig. 2 and Supplementary Table S2).

Fig. 1. Histogram of GO terms of protein coding genes for each species.

Gene families, selection and candidate genes

Total of 14 728 orthogroups (OrthoFinder) were found to contain *Impatiens* proteins, 77.5% of these also included one or more *Arabidopsis* proteins. A search for species specific orthology revealed a set of 236 unique groups that we consider as *Impatiens* specific. All species share 14 983 orthogroups, 5 246 orthogroups are missing in 1 species, 4 164 orthogroups are missing in 2 species, 6 803 orthogroups are missing in 3 species, 7 696 orthogroups were found in only 3 species, 11 518 orthogroups were found only in 2 species and 3 020 orthogroups were detected only in one species (Fig. 3, using OrthoFinder).

Fig. 3 Orthogroups overlap between all seven species. Grey boxes in lines with species names indicate presence of the given orthologous group in this species.

For each species, the CODEML program identified between 2 536 and 3 009 orthogroups under positive selection (Supplementary Table S3). These positively selected groups were assigned to GO categories. The highest proportions of mapped GO terms were related to binding (~28.2 %) and catalytic activity (~51.5%) under "Molecular Function", cellular (~31.2 %) and metabolic (35,3%) processes under "Biological Process", and cell part (~49.5%) and organelle (~35.5%) under "Cellular Component" (Fig. 4).

Fig. 4. Histogram of GO terms of the clusters of orthologous groups under positive selection for each species.

Also 767 orthogroups under positive selection common for all investigated species were identified. The highest proportions of mapped GO terms for these orthogroups were related to binding (25.2%) and catalytic activity (57.3%) under "Molecular Function", cellular (31.4%) and metabolic (37.3%) processes under "Biological Process", and cell part (52.8%) and organelle (31%) under "Cellular Component" (Fig. 5 and Supplementary Table S4).

Fig. 5 Histogram of GO terms of the clusters of orthologous groups under positive selection common for all species. Numbers represent number of identified orthologous genes for each category.

Discussion

We performed RNA sequencing of 7 closely related species of genus *Impatiens* to assess the differences and similarities in the transcriptome within the group. The results indicate good quality of assembled transcriptomes with comparable annotations across the species. From

283 identified orthologous gene families specific for *Impatiens* species, most abundant known
284 motives were reverse transcriptase motive, gag-polypeptide of long terminal repeat (LTR)
285 copia-type and Zinc knuckle motive. Six of the seven investigated species shared 77% of gene
286 families. Despite *I. bicornuta* distinct selection profile, this species shared selection on zing
287 finger protein structures and flowering regulation and stress response proteins with the other
288 investigated species. This consistency may suggest that the group evolved via adaptive
289 radiation.

290 Also similar transcriptome profile of *I. balsamina* compare to rest of the species was
291 identified, despite this species belongs, based on phylogeny Yu et al. (43), to the G clade, rest
292 of the species belong to clade B. This finding can be explained by study of Janssens et al.
293 (23). They hypothesized that rapid radiation of this genus is caused by a periodicity of glacial
294 cycles during the Pleistocene. That could have resulted in change of biotopes, with movement
295 of vegetation belts. To inhabit suitable localities, *Impatiens* had to migrate along with the
296 rainforest belts. We can expect, that climatic episodes isolated many different *Impatiens*
297 populations for several thousands of years. This explanation leads us to possible scenario that
298 species diversification is led by habitat requirements in given time in small isolated
299 populations. At present, these populations do not suffer from long term climate pressure, so
300 diversification and changes in stably transcribed genes can be slower, less pronounced, but
301 still running.
302 Natural individuals are rarely studied, but Baker et al. (20) studied transcriptomes of four pine
303 species. Some of the identified families are in common with our findings (F-box family,
304 ubiquitin related proteins, stress response family). Li et al. (63) compared more differentiated
305 species in *Rosaceae* (*Rosa, Malus, Prunus, Rubus,* and *Fragaria*) and they identified rapidly
306 evolving genes responsible for DNA damage repair, respond to environment stimuli and post
307 hybridization genome conflicts.

Characterization of de novo sequenced transcriptomes and between species transcriptome profiles comparison.

De novo transcriptome reconstruction

Despite necessary changes in RNA extraction protocol due to high concentration of secondary metabolites in the plant tissue, we were able to obtain good quality RNA extracts. Also high proportion (average 96%) of preserved reads after quality trimming, which is comparable with *I. balsamina* trimming (93%) in (40) indicates high quality of extracted RNA. Overall, RNA seq library construction with implemented changes was successful.

Comparison of the reads and subsequent assemblies showed many similarities in summary statistics among the sequenced transcriptomes. The biggest noticeable difference is in the size of the largest identified contigs. For three species, they are significantly smaller than in the other four. It can be caused by lower differences between isoforms, alternative splice forms, alleles, close paralogs, close homologs, and close homeologs compared to the other species. Six of the seven studied species have not been sequenced before. Only *I. balsamina* transcriptome has been recently sequenced (40). The transcriptome size we obtained for *I. balsamina* is similar to what has been previously published (84 635/ 91 873 transcripts). *Impatiens bicornuta* has the biggest detected transcriptome from all the investigated species (226 081 contigs) and also the longest assembly (196 656 333 bp). The smallest transcriptome and shortest assembly were identified for *I. scabrida* (94 707 contigs and 95 826 671 bp). Number of published reference transcriptomes for this genus is very limited. Transcriptome of *I. walleriana* (64) was also previously sequenced and its transcriptome size (121 497 contigs) is comparable with our results. As all our species were cultivated in identical conditions, the differences in transcriptome size between species can be caused by different habitats of plant origin and history of each species which cause changes in gene expression due to gene

expression regulation, epigenetic mechanisms, transposon activation, sequence changes, different combinations of regulatory factors and/or increased gene dosage. Nagano et al. (65) showed intra-annual fluctuation transcriptome size in *Arabidopsis halleri subsp. gemmifera*. They described from 2 879 to 7 185 seasonally oscillating genes. Gurung et al. (66) compared transcriptome size in a Himalaya plant (*Primula sikkimensis*) growing along an altitudinal gradient and carried out a transplant experiment within the altitudinal gradient. Total number of transcribed genes fluctuated between 38 423 and 48 674 depending on altitude of plant cultivation, which reflects species plasticity rather than adaptation.

The 86.7 % of contigs, which were correctly assembled corresponds to other assembled transcriptomes: 95.1% complete genes in (67), 91% in *Noccaea caerulescens* (68) and 94.4% in *Centaurium erythraea* (69). The highest (22%) proportion of missing and fragmented genes was identified for *I. scabrida.* This can be partially explained by mean length of the assembly, which is the highest (1021bp) for this species.

In terms of annotation, *I. devendrae* has the lowest percentage of undetermined transcripts (27.6%). Higher percentage of undetermined transcripts for the rest of the species is common for annotation without a reference genome (70). Higher number of unannotated transcripts can represent short or newly identified contigs which cannot be uniquely identified in public databases. We annotated more than 73 000 genes for each species, which is comparable with *I. walleriana* result (70 190). Our results are also comparable with other genera in *Ericales* order such as *Primula* (67 201 genes, (66)), *Phlox* (59 994 genes, (71)), *Saltugilia* (51 020-92 672 genes,(72)) , *Camellia* (46 223- 46 736, (73)) and *Vaccinium* (35 060 – 67 836 genes (74,75). GO terms assignment for identified transcripts is surprisingly similar between species. We expected that species differentiation and origin (altitudinal gradient) of sequenced individuals will affect overall transcriptome profiles. Our results suggest that species within this genus transcribe similar sets of genes and potential differences can be pronounced in

differential expression of these genes or their sequence. Similar results were published by Gurung et al. (66), who studied species plasticity by comparison of transcriptome size in *Primula sikkimensis* and they identified 21 167 transcripts and 109 and 85 genes which were differentially expressed between three altitudinal positions in which the plants were cultivated.

Orthologous gene families (Impatiens specific)

From identified orthologous gene families (14 728) for each species, we identified 334 gene families specific for *Impatiens* species. In these, 143 specific protein motives have been found. Fourteen of them were unknown, most abundant known motives were reverse transcriptase motive (7 times), gag-polypeptide of LTR copia-type (6) and Zinc knuckle motive (5). In group of 3 most common motives are integrase core domain, LTP family and RNA recognition motif. Majority of identified motives were found only once, so we will focus in this discussion on the most abundant hits only.

Reverse transcriptase, gag-polypeptide of LTR and integrase motive is expected in identified transcripts (76). These enzymes are generally connected with viruses and retrotransposons without specific function in plant cell and they are transcriptionally active. Lescot et al. (77) published that reverse transcriptase genes (RT) are more abundant in plankton metatrascriptome compare to a metagenome with distinct abundance patterns. They proposed that transcription of various RT-assisted elements could be involved in genome evolution or adaptive processes of plankton assemblage. Another possible source of transcripts was suggested by Gladyshev and Arkhipova (78) providing evidence of single copy reverse transcriptases. They found corresponding sequence in all major taxonomic groups including protists, fungi, animals, plants, and even bacteria evolving under strong selective pressure. Their function is not experimentally confirmed. Gag polypeptide motive is in congruence with

reverse transcription genes. They both are parts of retrotransposomes. These elements are studied in connection to environmental stress. Kalendar et al. (79) found correlation between LTR insertion in barley genome and ecogeographical distribution connected with BARE-1 promoter of abscisic acid-response elements typical for water stress-induced genes. Reverse transcriptase genes and LTRs opens interesting field for a molecular research to reveal their function and regulation in transcription level.

Zinc knuckle motive is studied from many perspectives. Loudet et al. (80) demonstrated that zinc knuckle protein negatively controls morning-specific growth in *Arabidopsis thaliana*. Also, proteins with this motive were found as a component of alternative splicing machinery in response to osmotic and salinity stress (81,82). Sequenced *Impatiens* species originated in different altitudes from Himalayas in Central and East Nepal. Osmotic or salinity stress is not expected environmental factor to shape transcription in these regions. More probably environmental factors as higher dosage of UV light and reactive oxygen species (83) will potentially change genome and/ or transcriptome response. Indeed, Zhao et al. (84) identified selection on zinc knuckle protein family as a part of genetic adaptation of *Lobelia aberdarica* and *L. telekii* transcriptomes to different altitudes in East African mountains as a response to DNA damage caused by volcanic eruptions, UV and frost damage. Zinc knuckle protein family (CX2CX3GHX4C) belongs to zinc finger superfamily (CCHC-type), so potential for functional diversity of this protein family is large.

Lipid transfer proteins (LTP family) and RNA recognition motives are involved in many processes, but two of them are common for both of them- pathogen defense and response to environmental stress (cold). Plants are continuously evolving complex regulatory networks to respond to pathogen threats (85) so specific *Impatiens* sequence changes are expected. Similarly, adaptation to cold in locations of origin of the investigated *Impatiens* species (1 330-2 728 m above sea level) is probably strong ecological driver of genome/transcriptome

406 changes. Specifically, lipid transfer proteins are involved in the intra- and extracellular
407 transport of lipids (86). Hincha et al. (87) published that member of LTP family, cryoprotectin
408 is involved as a protection of thylakoid membranes during a freeze-thaw cycle in cabbage.
409 Another member of LTP family (LTP 3) was also identified to be involved in response to
410 freezing in model species *A. thaliana* (88). Multiple proteins with glycine-rich RNA binding
411 proteins were found active during cold acclimation e.g. AtGRP7 (89), AtGR-RBP2 and
412 AtGR-RBP4 (90).

413 ## Gene families under positive selection and differences between species

414 Six of the seven investigated species shared 77% of gene families. Below, we on the 23% of
415 protein families differing among species and thus likely under selection and on distinct *I.*
416 *bicornuta* profile and their involvement in response to different environment stimuli. 50S
417 ribosome-binding GTPase domain is not discussed because lack of information.

418 **Resistance:** Resistance (R) proteins are involved in pathogen recognition and activation of
419 immune responses. Most resistance proteins contain a core nucleotide-binding domain. This
420 part is called NB-ARC domain and consists of three subdomains: NB, ARC1, and ARC2. The
421 NB-ARC domain was identified under selection only in *I. balsamina* transcriptome. It is a
422 functional ATPase domain, and its nucleotide-binding state regulates protein activity (91).
423 Recently, Ghelder et al. (92) reported fusion of an RPW8 (resistance to powdery mildew 8)
424 domain to a NB-ARC domain in conifers as a part of response to drought. Similarly, Bailey et
425 al. (93) performed phylogenetic analysis on plant immune receptors with NB-ARC domains
426 originated in grasses. They described that plant immune receptors are able to recognize
427 pathogen effectors through the acquisition of exogenous protein domains from other plant
428 genes through DNA transposition and/or ectopic recombination.

Arabidopsis broad-spectrum mildew resistance protein RPW8 is coded by naturally polymorphic locus and also belong to R proteins. Although genes are polymorphic, protein structure is stable (94). Zhong and Cheng (95) investigated RPW8 genes in 35 plant genomes and they found evidence for series of genetic events such as domain fissions, fusions, and duplications. Species-specific duplication events and tandemly duplicated clusters are processes responsible for species specific expansion.

Response to heavy metal stress: In *I. balsamina* heavy-metal-associated domain was found under selection. As name of the domain suggests, it is part of heavy-metal-associated proteins and they are involved in heavy metal detoxification. Li et al. (96) identified six clades in *A. thaliana* and *O. sativa* based on specificity of heavy metal-associated domains. They also identified tandem and segmental gene duplication and *cis*-acting elements on close proximity of promoter sequences. This finding suggests regulation by multiple transcription factors. Similar results were also published on *Populus trichocarpa* (97).

Protein families with zinc finger motive: B-box family was found under selection in two species: *I. scullyi* and *I. devendrae*. B-box family contains 32 zinc finger transcription factors which are involved in regulation of hormone signaling pathways and through them they regulate wide spectrum of processes such as drought-induced flowering pathway, germination, seedling photomorphogenesis, hypocotyl elongation in seedlings, induction of flowering, regulation of branching and shade avoidance response and many more. Despite conserved domain topology, this family is susceptible to environmental changes through variation in promotor region. In this region several different cis elements can be incorporated e.g. ABRE (ABA - Responsive element), ERE (EthyleneResponsive Element), CGTCA-motif and TGACG-motif (related to methyl-jasmonite stress responsiveness). Rosas et al. (98) identified natural variation in cis regulatory sequence of CONSTANS transcription factor

453 which regulates flowering time in *Arabidopsis*. They also showed recent evolution of
454 mutation and its spread to high frequency in *Arabidopsis* natural accessions.

455 Zn-finger in Ran binding protein family was detected in *I. balsamina, I. scullyi* and *I.*
456 *scabrida*. This protein family is very poorly studied. Part of proteins belonging to this family
457 is still uncharacterized. One of the characterized protein subfamilies in this family is
458 organellar zinc finger subfamily. These proteins are located in plastid and mitochondria. Their
459 function is not fully defined yet, but they are involved in chloroplast RNA editing (99).
460 Another example of Zn-finger motive in binding protein is histone deacetylase 15, which are
461 involved in the repression of chlorophyll synthesis in the dark (100). Also TATA-Binding
462 Protein-Associated Factor 15/15b belong to this group and is known to suppress flowering
463 during vernalization (101). More described in molecular functions is Ariadne subfamily in
464 Zn-finger family, a group of E3-type ubiquitin ligases, involved in last step of protein
465 degradation (102).

466 **DNA repair and protein translation**: RecA proteins were identified under selection in all
467 studied species except *I. scullyi*. Recombinases are responsible for homologous
468 recombination and maintenance of genome integrity. Recombinase RecA is crucial for DNA
469 repair. RecA bind to ssDNA break and scan for a homologous template dsDNA (103). Miller-
470 Messmer et al. (104) studied DNA recombination events in mitochondria. They found that
471 RecA-dependent repair has a dual effect on the mtDNA: maintaining the integrity of the
472 mitochondrial genome and also preferring the amplification of genome organization that
473 could help in the adaptation to environmental conditions.

474 CAAD domain of cyanobacterial aminoacyl-tRNA synthetase was found under selection in *I.*
475 *balsamina, I. racemosa* and *I. falcifer*: Aminoacyl-transfer RNA (tRNA) synthetases are key
476 molecules in translation and act early in protein synthesis by mediating the attachment of
477 amino acids to tRNA molecules. In plants, multiple versions of the protein are coded because

protein synthesis is running in three subcellular compartments (cytosol, mitochondria, and chloroplasts). Aminoacyl-transfer RNA (tRNA) synthetases are acquired in *Arabidopsis thaliana* genome through horizontal gene transfer event from bacteria. These genes in plant evolution were under selection to maintain original function, while additional copies diverged (105).

Regulation of flowering: CCT (chaperonin containing TCP-1) gene family was found under selection in *I. balsamina, I. racemosa, I. scabrida* and *I. falcifer*. Most CCT genes regulate flowering and they are studied mainly on rice for agricultural reasons (106) but studies on *Aegilops tauschii* proposed that CCT genes play important role in adaption of *A. tauschii* to the photoperiods of different regions. Zheng et al. (107) identified rapid evolution of CCT genes. More gene variations are available for adaption to environmental changes, so positive selection has kept more convenient variations.

Nitrate detection and metabolism: Fip1 motive was found under selection in *I. balsamina, I. racemosa* and *I. falcifer*. Fip1 gene plays important role in nitrate signaling through expression regulates of kinesis CIPK8 and CIPK2 in *Arabidopsis* (108). Nitrate content in environment is very important indicator and regulatory networks consist of approximately 300 genes differently regulated under variety of conditions (109). These findings showed that nitrate signaling is under environmental pressure and adaptability to nitrate uptake, transport, assimilation and metabolism is crucial.

Cell wall and membrane polymers: N-acetylglucosamine transporters were identified under selection in *I. balsamina, I. scullyi* and *I. falcifer*. The biosynthesis of glycans, glycoproteins and glycolipids requires glycosyltransferases localized in the Golgi apparatus and endoplasmic reticulum (110). They were identified as overexpressed in rice after salt stress exposure (111) and are also required for an arbuscular mycorrhizal presymbiotic fungal reprogramming (112).

Families with pleitropic effects: Ras superfamily domain and GTPase domains were identified under selection in all investigated species. Ras superfamily consists of Rab, Ras, Arf, Rho and Ran families in yeast and animals, but the Ras family has not been found in plants. This superfamily is one of the most important gene families involved in signal transduction, vesicle trafficking, signaling, cytoskeleton rearrangements, nuclear transport, cell growth, plant defense in interaction with microorganisms (113). Their important role is also in adaptation to oxygen deprivation through levels of H_2O_2. H_2O_2 acts as a second messenger for either activate expression of genes responsible for tolerance against oxygen deprivation (e.g. the gene encoding alcohol dehydrogenase) or activate the generation of intracellular Ca^{2+} signals responsible for cell growth (114).

Specific selection profile of *I. bicornuta*

Stress response: AP180 N-terminal homology (ANTH) domain is studied just recently and their functions are not fully understood yet. *A. thaliana* genome contains 18 ANTH proteins which showed their possible diversification in comparison with fewer ANTH proteins in metazoan and fungal genomes (115). Nguyen et al. (116) showed that mutant in ANTH protein, ateca4 showed higher resistance to osmotic stress and more sensitivity to exogenous abscisic acid molecules.

Hsp20 domain belongs to small heat shock proteins that are regulated by stress and work as molecular chaperons to protect proteins from stress related damage (117). Some of them evolved only recently, others are ancient (118). This region is conserved in C-terminal, but in 2 subfamilies β6 sheet is missing in plants and some animals. This motive is necessary for dimerization and oligomerization (119), but investigated proteins are able to create dimers anyway (120). On the other hand, N-terminal domain is very variable. Despite unstable

structure, this region is functionally important because of substrate binding specificity (121). It is possible to conclude that these proteins have capacity for diversification.

Stress responsive A/B Barrel Domain forms dimers and its function is not fully understood. Shaik and Ramakrishna (122) found this domain/protein co-expressed under drought and bacterial stresses in *A. thaliana* and *Oryza sativa*. Protein with this domain was also identified downregulated in salt stressed *Abelmoschus esculentus* L. seedlings (123).

Uroporphyrinogen decarboxylase (URO-D) is involved in chlorophyll biosynthesis (UniProtKB - Q93ZB6). Mock et al. (124) identified this enzyme as a part of plant defense respond to stress caused by reactive oxygen species in *Nicotinia tabacum*. Vanhove et al. (125) also identified this protein significantly abundant as response to drought stress in banana. Zhu et al. (126) found increased abundance in this protein in waterlogged *Vitis vinifera*.

Cellular processes: Components of kinetochore, centromere proteins A are proteins under positive selection in plants. They contain DNA binding motive. Centromeres are built from repetitive satellite sequences that are rapidly evolving, so proteins with binding specificity must evolve accordingly (127).

F-box proteins control protein ubiquitination by ubiquitin 26S-proteasome system. They are involved in wide spectrum of processes including secondary metabolite regulation, response to stresses, phytohormone signaling, developmental processes and miRNA biogenesis (128). Schumann et al. (129) did phylogeny analysis of largest F-box subfamily because in non-plant eukaryotes genomes only one copy is present compared to *A. thaliana* with 103 copies in genome. They identified signatures of positive selection and part of the diverse protein domain which is lineage specific. These findings showed adaptation potential of this family.

549 Maintenance of mitochondrial morphology protein 1 is very poorly studied. This protein is
550 part of protein complex, which physically connect endoplasmatic reticulum and mitochondria.
551 Wang et al. (130) showed upregulation of this gene in *Aspergillus nidulans* under salinity
552 stress.
553 RimM (ribosome maturation factor) N-terminal domain belongs to RimM protein, which is
554 associated with 30S ribosomal subunit. This indicates its function in translation initialization
555 or subunit maturation (131). It is been found in some genomes (*Plasmodium falciparum* and
556 *P. yoelii*, *Anopheles gambiae*, *A. thaliana* (132)) but very little is known about it.

557 **Protein families with zinc finger motive:** FAR1 (FAR-RED IMPAIRED RESPONSE1)
558 DNA-binding domain is an N-terminal C2H2 zinc-finger domain. It is part of well
559 characterized transposase-derived transcription factor FAR1. This transcription factor is
560 involved in many processes e.g. light signal transduction, chloroplast division, chlorophyll,
561 myo- inositol and starch biosynthesis, circadial clock regulation and drought and low
562 phosphate response (133). Salojärvi et al. (134) showed that this gene family can be involved
563 in adaptation to environment in *Betula pendula*.

564 Zinc-finger of the FCS-type (phenyl alanine and serine residues associated with the third
565 cysteine) form one of the largest transcription factor families in plants. They are categorized
566 into subfamilies based on the order of the Cys and His residues in their secondary structures
567 (135). Such a large family is involved in many stress related pathways. They have a role in
568 salt, cold, osmotic, drought, oxidative and high- light stress. Specifically, Jamsheer et al (136)
569 showed that downregulation of FCS-LIKE ZINC FINGER 6 and 10 is part of osmotic stress
570 responses in *A. thaliana*. Emerson and Thomas (137) published analysis of the gene family
571 across the tree of life and identified positive selection acting on this family and specifically
572 selection to change DNA-binding specificity of transcription factors.

Family with pleitropic effects: Homeobox domain is present in 14 classes of transcription factors in plants. They are involved in wide spectrum of processes (signal transduction networks involved in response to abscisic acid, auxin and drought, general growth regulation, in shoot and floral meristem development, stem-cell specification and proliferation, flowering time regulation directly repress gibberellin and lignin biosynthetic gene expression and in the adaptation to drought and disease resistance to fungal pathogens and many more (138). More recently Khan et al. (139) investigated involvement of homeobox genes in *Brassica rapa* under various stress conditions. They identified active purifying selection on this gene family and dynamic variations in differential expression combined with their responses against multiple stresses.

Energy metabolism: Oxidoreductase FAD-binding domain is part of flavoprotein oxidoreductases molecule. They act in oxidative pathways and are able to oxidize NADH/ $FADH_2$. These molecules are well studied in different organisms but recently Trisolini et al. (140) compared sequences and crystal structures across selected taxonomic groups (*Bacillales* , *Enterobacteriales*, *Rhodospirillales* , *Rhodobacterales* , *Thermales* ,*Rhizobia, Nematoda*, *Mammalia* , *Arthropoda*, *Anthozoa*, *Fungi*, *Plants and H. sapiens*, *S. cerevisiae*, *C. thermarum*). They showed that these proteins have different coding sequences, but despite that, protein structure and shapes are similar and functional in oxidation processes.

Flowering regulation: SNW domain in SKIP protein regulate pre-mRNA splicing and also was identified as a transcription factor of gene transcription activator in Paf1 complex, which is involved in transcription of FLOWERING LOCUS C and flowering in *A. thaliana* (10). It is involved in alternative splicing regulation of clock and salt tolerance-related genes in plants (141).

Conclusions

We performed RNA sequencing of 7 closely related species of genus *Impatiens*. Most abundant motives were reverse transcriptase motive, gag-polypeptide of LTR copia-type and Zinc knuckle motive from orthologous, *Impatiens* specific gene families. Six of the seven investigated species shared 77% of gene families under selection. We also described distinct *I. bicornuta* selection profile, but this species shared selection on zing finger protein structures and flowering regulation and stress response proteins with the other investigated species. The reason for the strong differentiation of this species, however, remains unclear, but shared selection may indicate evolution via adaptive radiation in *Impatiens* species.

It is possible to conclude that genes involved in response to environmental stimuli is a major driver for selection in *Impatiens* species. More natural populations and species have to be sequenced to reveal conserved biological principles and distinguish lineage and locality adaptations.

Acknowledgements

We thank to P. Vácha (Seqme) for excellent sequencing help, M. Rokaya for collection of the seeds used for the study and Z. Líblová for cultivation of the plants.

References

1. Frodin DG. History and concepts of big plant genera. Taxon. 2004;53(3):753–76.

2. Schluter D. The ecology of adaptive radiation. OUP Oxford. 2000.

3. Pincheira-Donoso D, Harvey LP, Ruta M. What defines an adaptive radiation? Macroevolutionary diversification dynamics of an exceptionally species-rich continental lizard radiation. BMC Evol Biol. 2015;15(1):1–13.

620 4. Venditti C, Meade A, Pagel M. Phylogenies reveal new interpretation of speciation and
621 the Red Queen. Nature. 2010;463(7279):349–52.

622 5. Alix K, Gérard PR, Schwarzacher T, Heslop-Harrison JS. Polyploidy and interspecific
623 hybridization: partners for adaptation, speciation and evolution in plants. Ann Bot.
624 2017;120(2):183–94.

625 6. Solbrig OT. Fertility, sterility and the species problem. Fertil Steril species Probl. 1968;

626 7. Cullis CA. DNA rearrangements in response to environmental stress. Adv Genet.
627 1990;28:73–97.

628 8. Fiebig A, Kimport R, Preuss D. Comparisons of pollen coat genes across Brassicaceae
629 species reveal rapid evolution by repeat expansion and diversification. Proc Natl Acad
630 Sci. 2004;101(9):3286–91.

631 9. Bromham L, Hua X, Lanfear R, Cowman PF. Exploring the relationships between
632 mutation rates, life history, genome size, environment, and species richness in
633 flowering plants. Am Nat. 2015;185(4):507–24.

634 10. Cao Y, Wen L, Wang Z, Ma L. SKIP Interacts with the Paf1 Complex to Regulate
635 Flowering via the Activation of FLC Transcription in Arabidopsis. Mol Plant.
636 2015;8(12):1816–9.

637 11. Bagheri A, Maassoumi AA, Rahiminejad MR, Brassac J, Blattner FR. Molecular
638 phylogeny and divergence times of Astragalus section Hymenostegis: An analysis of a
639 rapidly diversifying species group in Fabaceae. Sci Rep. 2017;7(1):1–9.

640 12. Dorsey B. Phylogenetics and Morphological Evolution of Euphorbia subgenus
641 Euphorbia. 2013.

642 13. Nepokroeff M, Bremer B, Sytsma KJ. Reorganization of the genus Psychotria and tribe

643 Psychotrieae (Rubiaceae) inferred from ITS and rbcL sequence data. Syst Bot. 1999;5–
644 27.

645 14. Veeckman E, Ruttink T, Vandepoele K. Are we there yet? Reliably estimating the
646 completeness of plant genome sequences. Plant Cell. 2016;28(8):1759–68.

647 15. Schreiber AW, Sutton T, Caldo RA, Kalashyan E, Lovell B, Mayo G, et al.
648 Comparative transcriptomics in the Triticeae. BMC Genomics. 2009;10(1):1–17.

649 16. Amrine KCH, Blanco-Ulate B, Riaz S, Pap D, Jones L, Figueroa-Balderas R, et al.
650 Comparative transcriptomics of Central Asian Vitis vinifera accessions reveals distinct
651 defense strategies against powdery mildew. Hortic Res. 2015;2(1):1–11.

652 17. Qiao Q, Xue L, Wang Q, Sun H, Zhong Y, Huang J, et al. Comparative transcriptomics
653 of strawberries (Fragaria spp.) provides insights into evolutionary patterns. Front Plant
654 Sci. 2016;7(1839).

655 18. Zhu S, Tang S, Tan Z, Yu Y, Dai Q, Liu T. Comparative transcriptomics provide
656 insight into the morphogenesis and evolution of fistular leaves in Allium. BMC
657 Genomics. 2017;18(1):1–9.

658 19. Shimizu-Inatsugi R, Terada A, Hirose K, Kudoh H, Sese J, Shimizu KK. Plant adaptive
659 radiation mediated by polyploid plasticity in transcriptomes. Mol Ecol.
660 2017;26(1):193–207.

661 20. Baker EAG, Wegrzyn JL, Sezen UU, Falk T, Maloney PE, Vogler DR, et al.
662 Comparative transcriptomics among four white pine species. G3 Genes, Genomes,
663 Genet. 2018;8(5):1461–74.

664 21. Christenhusz MJM, Byng JW. The number of known plants species in the world and its
665 annual increase. Phytotaxa. 2016;261(3):201–17.

666 22. Grey-Wilson C. Impatiens of Africa. CRC Press. 1980.

667 23. Janssens SB, Knox EB, Huysmans S, Smets EF, Merckx VSFT. Rapid radiation of
668 Impatiens (Balsaminaceae) during Pliocene and Pleistocene: Result of a global climate
669 change. Mol Phylogenet Evol. 2009;52(3):806–24.

670 24. Kaufman W, Kaufman SR. Invasive plants: guide to identification and the impacts and
671 control of common North American species. Stackpole Books; 2013.

672 25. Donohue K, Messiqua D, Pyle EH, Shane Hesche M, Schmitt J. Evidence of adaptive
673 divergence in plasticity: Density- and site-dependent selection on shade-avoidance
674 responses in impatiens capensis. Evolution. 2000;54(6):1956–68.

675 26. Dostálek T, Bahadur Rokaya M, Münzbergová Z. Plant palatability and trait responses
676 to experimental warming. Sci Rep. 2020;10(1):1–12.

677 27. Gruntman M, Segev U, Tielbörger K. Shade-induced plasticity in invasive Impatiens
678 glandulifera populations. Weed Res. 2020;60(1):16–25.

679 28. Hattori M, Nagano Y, Shinohara Y, Itino T. Pattern of flower size variation along an
680 altitudinal gradient differs between Impatiens textori and Impatiens noli-tangere. J
681 Plant Interact. 2016;11(1):152–7.

682 29. Masuda M, Yahara T, Maki M. Evolution of floral dimorphism in a cleistogamous
683 annual, Impatiens noli-tangere L. occurring under different environmental conditions.
684 Ecol Res. 2004;19(6):571–80.

685 30. Mitchell-Olds T, Bergelson J. Statistical genetics of an annual plant, Impatiens
686 capensis. II. Natural selection. Genetics. 1990;124(2):417–21.

687 31. Münzbergová Z, Kosová V, Schnáblová R, Rokaya M et al. Plant origin, but not
688 phylogeny, drive species ecophysiological response to projected climate.

2020;11(400):1–18.

32. Schoen DJ, Latta RG. Spatial autocorrelation of genotypes in populations of impatiens pallida and impatiens capensis. Heredity. 1989;63(2):181–9.

33. Veselá A, Dostálek T, Rokaya MB, Münzbergová Z. Seed mass and plant home site environment interact to determine alpine species germination patterns along an elevation gradient. Alp Bot. 2020;130(2):101–13.

34. Luo C, Huang W, Li Y, Feng Z, Zhu J, Liu Y, et al. The complete chloroplast genome sequence of horticultural plant, Impatiens hawkeri (Sect. Balsaminacea, Impatiens). Mitochondrial DNA Part B. 2020;5(1):119–20.

35. Luo C, Huang W, Zhu J, Feng Z, Liu Y, Li Y, et al. The complete chloroplast genome of Impatiens uliginosa Franch., an endemic species in Southwest China. Mitochondrial DNA Part B Resour. 2019;4(2):3846-3847.

36. Li ZZ, Saina JK, Gichira AW, Kyalo CM, Wang QF, Chen JM. Comparative genomics of the balsaminaceae sister genera Hydrocera triflora and Impatiens pinfanensis. Int J Mol Sci. 2018;19(1):319.

37. Li Q, Li X, Zhuo M, Qieyang R, Dongzhi D. Characterization of the complete chloroplast genome of Impatiens alpicola (Balsaminaceae : Impatiens), a rare and endemic Chinese flowering plant. Mitochondrial DNA Part B. 2019;4(2):3646–7.

38. Wang Q, Li WQ, Ding B, Deng HP. Characterization of the complete chloroplast genome sequence of Impatiens pritzelii (Balsaminaceae): an endemic species from China. Mitochondrial DNA Part B Resour. 2019;4(2):4073–4.

39. Kurose D, Pollard KM, Ellison CA. Chloroplast DNA analysis of the invasive weed , Himalayan balsam (Impatiens glandulifera), in the British Isles. Sci Rep.

2020;10(1):1–12.

40. Foong LC, Ho ASH, Yeo BPH, Lim YM, Tam SM. Data of de novo assembly and functional annotation of the leaf transcriptome of Impatiens balsamina. Data Br. 2019;23(103603).

41. Akiyama S, Ohba H. Studies of impatiens (Balsaminaceae) of Nepal 3. Impatiens scabrida and Allied species. Bull Natl Mus Nat Sci Ser B Bot. 2016;42:121–30.

42. Fisher E. Balsaminaceae. In: Flowering plants Dicotyledons: Celastrales, Oxalidales, Rosales, Cornales, Ericales Vol6. Springer Science & Business Media; 2013.

43. Yu S, Janssens SB, Zhu X, Lid M. Cladistics Phylogeny of Impatiens (Balsaminaceae): integrating molecular and morphological evidence into a new classification. Cladistics. 2016;32(2):179–97.

44. Yuan Y-M. Phylogeny and biogeography of Balsaminaceae inferred from ITS sequences. Taxon. 2004;53(2):391–404.

45. Foong L, Ho S, Lim Y, Tam S. A modified CTAB-based protocol for total RNA extraction from the medicinal plant Impatiens balsamina (Balsaminaceae) for next-generation sequencing studies. Malaysian Appl Biol J. 2017;46(2):87–95.

46. MacManes MD. Establishing evidenced-based best practice for the de novo assembly and evaluation of transcriptomes from non-model organisms. bioRxiv. 2015;035642.

47. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

48. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

734 49. Song L, Florea L. Rcorrector: Efficient and accurate error correction for Illumina
735 RNA-seq reads. Gigascience. 2015;4(1):s13742-015.

736 50. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
737 transcriptome assembly from RNA-Seq data without a reference genome. Nat
738 Biotechnol. 2011;29(7):644.

739 51. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: Quality assessment tool for
740 genome assemblies. Bioinformatics. 2013;29(8):1072–5.

741 52. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:
742 Assessing genome assembly and annotation completeness with single-copy orthologs.
743 Bioinformatics. 2015;31(19):3210–2.

744 53. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de
745 novo transcriptome assemblies from RNA-Seq data. Genome Biol. 2014;15(12):1–21.

746 54. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De
747 novo transcript sequence reconstruction from RNA-seq using the Trinity platform for
748 reference generation and analysis. Nat Protoc. 2013;8(8):1494–512.

749 55. Finn RD, Clements J, Eddy SR. HMMER web server: Interactive sequence similarity
750 searching. Nucleic Acids Res. 2011;39(suppl_2):W29–37.

751 56. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer:
752 Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res.
753 2007;35(9):3100–8.

754 57. Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. PANTHER version 10:
755 Expanded protein families and functions, and analysis tools. Nucleic Acids Res.
756 2016;44(D1):D336–42.

757 58. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
758 comparisons dramatically improves orthogroup inference accuracy. Genome Biol.
759 2015;16(1):1–14.

760 59. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
761 Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.

762 60. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence
763 alignments into the corresponding codon alignments. Nucleic Acids Res.
764 2006;34(suppl_2):W609–12.

765 61. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of
766 large phylogenies. Bioinformatics. 2014;30(9):1312–3.

767 62. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol.
768 2007;24(8):1586–91.

769 63. Li S, Zhong M, Dong X, Jiang X, Xu Y, Sun Y, et al. Comparative transcriptomics
770 identifies patterns of selection in roses. BMC Plant Biol. 2018;18(1):1–12.

771 64. Suarez S, Naveed ZA, Ali GS. Transcriptional profiling of Impatiens walleriana genes
772 through different stages of downy mildew infection reveals novel genes involved in
773 disease susceptibility. bioRxiv. 2019;622480.

774 65. Nagano AJ, Kawagoe T, Sugisaka J, Honjo MN, Iwayama K, Kudoh H. Annual
775 transcriptome dynamics in natural environments reveals plant seasonal adaptation. Nat
776 Plants. 2019;5(1):74–83.

777 66. Gurung PD, Upadhyay AK, Bhardwaj PK, Sowdhamini R, Ramakrishnan U.
778 Transcriptome analysis reveals plasticity in gene regulation due to environmental cues
779 in Primula sikkimensis, a high altitude plant species. BMC Genomics. 2019;20(1):1–

780 12.

781 67. Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J, et al. High quality de
782 novo transcriptome assembly of Croton tiglium. Front Mol Biosci. 2018;5(62):1–5.

783 68. Blande D, Halimaa P, Tervahauta AI, Aarts MGM, Kärenlampi SO. De novo
784 transcriptome assemblies of four accessions of the metal hyperaccumulator plant
785 Noccaea caerulescens. Sci data. 2017;4(1):1–9.

786 69. Ćuković K, Dragićević M, Bogdanović M, Paunović D, Giurato G, Filipović B, et al.
787 Plant regeneration in leaf culture of Centaurium erythraea Rafn. Part 3: de novo
788 transcriptome assembly and validation of housekeeping genes for studies of in vitro
789 morphogenesis. Plant Cell Tissue Organ Cult. 2020;141(2):417–33.

790 70. Dombrowski JE, Kronmiller BA, Hollenbeck VG, Rhodes AC, Henning JA, Martin
791 RC. Transcriptome analysis of the model grass Lolium temulentum exposed to green
792 leaf volatiles. BMC Plant Biol. 2019;19(1):1–17.

793 71. Qu Y, Zhou A, Zhang X, Tang H, Liang M, Han H, et al. De novo transcriptome
794 sequencing of low temperature-treated Phlox subulata and analysis of the genes
795 involved in cold stress. Int J Mol Sci. 2015;16(5):9732–48.

796 72. Landis JB, Soltis DE, Soltis PS. Comparative transcriptomic analysis of the evolution
797 and development of flower size in Saltugilia (Polemoniaceae). BMC Genomics.
798 2017;18(1):1–15.

799 73. Li Y, Wang X, Ban Q, Zhu X, Jiang C, Wei C, et al. Comparative transcriptomic
800 analysis reveals gene expression associated with cold adaptation in the tea plant
801 Camellia sinensis. BMC Genomics. 2019;20(1):1–17.

802 74. Zhang F, Ji S, Wei B, Cheng S, Wang Y, Hao J, et al. Transcriptome analysis of

postharvest blueberries (Vaccinium corymbosum 'Duke') in response to cold stress. BMC Plant Biol. 2020;20(1):1–20.

75. Tian Y, Ma Z, Ma H, Gu Y, Li Y, Sun H. Comparative transcriptome analysis of lingonberry (Vaccinium vitis-idaea) provides insights into genes associated with flavonoids metabolism during fruit development. Biotechnol Biotechnol Equip. 2020;34(1):1252–64.

76. Tokuyama M, Kong Y, Song E, Jayewickreme T, Kang I, Iwasaki A. ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses. Proc Natl Acad Sci. 2018;115(50):12565–72.

77. Lescot M, Hingamp P, Kojima KK, Villar E, Romac S, Veluchamy A, et al. Reverse transcriptase genes are highly abundant and transcriptionally active in marine plankton assemblages. ISME J. 2015;10(5):1134–46.

78. Gladyshev EA, Arkhipova IR. A widespread class of reverse transcriptase-related cellular genes. Proc Natl Acad Sci. 2011;108(51):20311–6.

79. Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci. 2000;97(12):6603–7.

80. Loudet O, Michael TP, Burger BT, Mette C Le, Weigel D, Chory J. A zinc knuckle protein that negatively controls morning-specific growth in Arabidopsis thaliana. Proc Natl Acad Sci. 2008;105(44):17193–8.

81. Palusa SG, Ali GS, Reddy ASN. Alternative splicing of pre-mRNAs of Arabidopsis serine / arginine-rich proteins : regulation by hormones and stresses. Plant J. 2007;49(6):1091–107.

826 82. Ding F, Cui P, Wang Z, Zhang S, Ali S, Xiong L. Genome-wide analysis of alternative
827 splicing of pre-mRNA under salt stress in Arabidopsis. BMC Genomics. 2014;15(1):1–
828 14.

829 83. Sedej TT, Erznožnik T, Rovtar J. Effect of UV radiation and altitude characteristics on
830 the functional traits and leaf optical properties in Saxifraga hostii at the alpine and
831 montane sites. Photochem Photobiol Sci. 2019;19(2):180–92.

832 84. Zhao S, Chen L, Muchuku JK, Hu G, Wang Q. Genetic Adaptation of Giant Lobelias (
833 Lobelia aberdaric a and Lobelia telekii) to Different Altitudes in East African
834 Mountains. Front Plant Sci. 2016;7(488):1–9.

835 85. Imran QM, Yun B. Pathogen-induced Defense Strategies in Plants. J Crop Sci
836 Biotechnol. 2020;23(2):97–105.

837 86. Finkina EI, Melnikova DN, Bogdanov I V, Ovchinnikova T V. Lipid Transfer Proteins
838 As Components of the Plant Innate Immune System: Structure, Functions and
839 Applications. Acta Naturae. 2016;8(29):47–61.

840 87. Hincha DK, Neukamm B, Sror HAM, Sieg F, Weckwarth W, Schro W. Cabbage
841 cryoprotectin is a member of the nonspecific plant lipid transfer protein gene family.
842 Plant Physiol. 2001;125(2):835–46.

843 88. Guo L, Yang H, Zhang X, Yang S. Lipid transfer protein 3 as a target of MYB96
844 mediates freezing and drought stress in Arabidopsis. J Exp Bot. 2013;64(6):1755–67.

845 89. Kim JS, Park SJ, Kwak KJ, Kim YO, Kim JY, Song J, et al. Cold shock domain
846 proteins and glycine-rich RNA-binding proteins from Arabidopsis thaliana can promote
847 the cold adaptation process in Escherichia coli. Nucleic Acids Res. 2007;35(2):506–16.

848 90. Vermel M, Guermann B, Delage L, Grienenberger J, Mare L. A family of RRM-type

RNA-binding proteins specific to plant mitochondria. Proc Natl Acad Sci. 2002;99(9):5866–71.

91. Van Ooijen G, Mayr G, Kasiem MMA, Albrecht M, Cornelissen BJC, Takken FLW. Structure–function analysis of the NB-ARC domain of plant disease resistance proteins. J Exp Bot. 2008;59(6):1383–97.

92. Ghelder C Van, Parent GJ, Rigault P, Prunier J, Giguère I, Caron S, et al. The large repertoire of conifer NLR resistance genes includes drought responsive and highly diversified RNLs. Sci Rep. 2019;9(1):1–13.

93. Bailey PC, Schudoma C, Jackson W, Baggs E, Dagdas G, Haerty W, et al. Dominant integration locus drives continuous diversification of plant immune receptors with exogenous domain fusions. Genome Biol. 2018;19(1):1–18.

94. B S, Emerson B, Ratanasut K, Patrick E, Neill CO, Bancroft I, et al. Origin and maintenance of a broad-spectrum disease resistance locus in Arabidopsis. Mol Biol Evol. 2004;21(9):1661–72.

95. Zhong Y, Cheng ZM. A unique RPW8-encoding class of genes that originated in early land plants and evolved through domain fission, fusion and duplication. Sci Rep. 2016;6(1):1–13.

96. Li J, Zhang M, Sun J, Mao X, Wang J, Liu H. Heavy metal stress-associated proteins in rice and Arabidopsis : Genome-wide identification, phylogenetics, duplication , and expression profiles analysis. Front Genet. 2020;11:1–21.

97. Li D, Xu X, Hu X, Liu Q, Wang Z. Genome-wide analysis and heavy metal-induced expression profiling of the HMA gene family in Populus trichocarpa. Front Plant Sci. 2015;6(1149).

872 98. Rosas U, Mei Y, Xie Q, Banta JA, Zhou RW, Seufferheld G, et al. Variation in
873 Arabidopsis flowering time associated with cis -regulatory variation in CONSTANS.
874 Nat Commun. 2014;5(1):1–8.

875 99. Sun T, Shi X, Friso G, Wijk K Van, Bentolila S, Hanson MR. A zinc finger motif-
876 containing protein is essential for chloroplast RNA editing. PLoS Genet.
877 2015;11(3):e1005028.

878 100. Liu X, Chen C, Wang K, Luo M, Tai R, Yuan L, et al. PHYTOCHROME
879 INTERACTING FACTOR3 associates with the histone deacetylase HDA15 in
880 repression of chlorophyll biosynthesis and photosynthesis in etiolated Arabidopsis
881 seedlings. Plant Cell. 2013;25(4):1258–73.

882 101. Eom H, Park SJ, Kim MK, Kim H, Kang H, Lee I. TAF15b, involved in the
883 autonomous pathway for flowering, represses transcription of FLOWERING LOCUS
884 C. Plant J. 2018;93(1):79–91.

885 102. Hershko A, Ciechanover A. THE UBIQUITIN SYSTEM. Annu Rev Biochem.
886 1998;67(1):425–79.

887 103. Del Val E, Nasser W, Abaibou H, Reverchon S. RecA and DNA recombination: a
888 review of molecular mechanisms. Biochem Soc Trans. 2019;47(5):1511–31.

889 104. Miller-Messmer M, Kühn K, Bichara M, Le Ret M, Imbault P, Gualberto JM. RecA-
890 dependent DNA repair results in increased heteroplasmy of the Arabidopsis. Plant
891 Physiol. 2012;159(1):211–26.

892 105. Brandao MM, Silva-Filho MC. Evolutionary history of Arabidopsis thaliana
893 aminoacyl-tRNA synthetase dual-targeted proteins. Mol Biol Evol. 2011;28(1):79–85.

894 106. Zhang J, Hu Y, Xu L-H, He Q, Fan X, Xing Y. The CCT domain-containing gene

family has large impacts on heading date, regional adaptation and grain yield in rice. J Integr Agric. 2017;16(12):2686–97.

107. Zheng X, Li X, Ge C, Chang J, Shi M, Chen J, et al. Characterization of the CCT family and analysis of gene expression in Aegilops tauschii. PLoS One. 2017;12(12):e0189333.

108. Wang C, Zhang W, Li Z, Li Z, Bi Y, Crawford NM, et al. FIP1 plays an important role in nitrate signaling and regulates CIPK8 and CIPK23 expression in Arabidopsis. Front Plant Sci. 2018;9(593).

109. Gutiérrez RA, Gifford ML, Poultney C, Wang R, Shasha DE, Gutie RA, et al. Insights into the genomic nitrate response using genetics and the Sungear Software System. J Exp Bot. 2007;58(9):2359–67.

110. Niemann MCE, Bartrina I, Ashikov A, Weber H. Arabidopsis ROCK1 transports UDP-GlcNAc/UDP-GalNAc and regulates ER protein quality control and cytokinin activity. Proc Natl Acad Sci. 2015;112(1):291–6.

111. Razzaq A, Ali A, Safdar L Bin, Zafar MM, Rui Y, Shakeel A, et al. Salt stress induces physiochemical alterations in rice grain composition and quality. J Food Sci. 2020;85(1):14–20.

112. Nadal M, Sawers R, Naseem S, Bassin B, Kulicke C, Sharman A, et al. An N-acetylglucosamine transporter required for arbuscular mycorrhizal symbioses in rice and maize. Nat plants. 2017;3(6):1–7.

113. Rivero C, Traubenik S. Small GTPases in plant biotic interactions. Small GTPases. 2019;10(5):350–60.

114. Gu Y, Wang Z, Yang Z. ROP / RAC GTPase : an old new master regulator for plant

signaling. 2004;7(5):527-536.

115. Muro K, Matsuura-tokita K, Tsukamoto R, Kanaoka MM, Ebine K, Higashiyama T, et al. ANTH domain-containing proteins are required for the pollen tube plasma membrane integrity via recycling ANXUR kinases. Commun Biol. 2018;1(1):1–9.

116. Nguyen HH, Lee MH, Song K, Ahn G, Lee J, Hwang I. The A/ENTH domain-containing protein AtECA4 is an adaptor protein involved in cargo recycling from the trans-Golgi network/early endosome to the plasma membrane. Mol Plant. 2018;11(4):568–83.

117. Waters ER, Vierling E. Plant small heat shock proteins – evolutionary and functional diversity. Plant small heat Shock proteins–evolutionary Funct Divers. 2020;227(1):24–37.

118. Waters ER, Aevermann BD, Sanders-Reed Z. Comparative analysis of the small heat shock proteins in three angiosperm genomes identifies new subfamilies and reveals diverse evolutionary patterns. Cell Stress Chaperones. 2008;13(2):127–42.

119. Montfort BROBVAN, Slingsby C, Elizabeth VIERLING. Structure and function of the small heat shock protein/α-crystallin family of molecular chaperones. Adv Protein Chem. 2001;59:105–56.

120. Siddique M, Gernhard S. The plant sHSP superfamily : five new members in Arabidopsis thaliana with unexpected properties. Cell Stress Chaperones. 2008;13(2):183–97.

121. McDonald ET, Bortolus M, Koteiche HA, Mchaourab HS. Sequence, structure and dynamic determinants of Hsp27 (HspB1) equilibrium dissociation are encoded by the N-terminal Domain. Biochemistry. 2013;51(6):1257–68.

941 122. Shaik R, Ramakrishna W. Genes and Co-Expression Modules Common to Drought and
942 Bacterial Stress Responses in Arabidopsis and Rice. 2013;8(10):1–16.

943 123. Zhan Y, Wu Q, Chen Y, Tang M, Sun C, Sun J, et al. Comparative proteomic analysis
944 of okra (Abelmoschus esculentus L .) seedlings under salt stress. BMC Genomics.
945 2019;20(1):1–12.

946 124. Mock H, Keetman U, Kruse E, Rank B, Grimm B. Defense responses to tetrapyrrole-
947 induced oxidative stress in transgenic plants with reduced uroporphyrinogen
948 decarboxylase or coproporphyrinogen oxidase activity. Plant Physiol.
949 1998;116(1):107–16.

950 125. Vanhove A, Vermaelen W, Panis B, Swennen R. Screening the banana biodiversity for
951 drought tolerance : can an in vitro growth model and proteomics be used as a tool to
952 discover tolerant varieties and understand homeostasis. Front Plant Sci. 2012;3(176).

953 126. Zhu X, Li X, Jiu S, Zhang K, Wang C, Fang J. Analysis of the regulation networks in
954 grapevine reveals response to waterlogging stress and candidate gene-marker selection
955 for damage severity. R Soc open Sci. 2018;5(6):172253.

956 127. Talbert PB, Bryson TD, Henikoff S. Adaptive evolution of centromere proteins in
957 plants and animals. J Biol. 2004;3(4):1–17.

958 128. Athirah N, Hamid A, Izzat M, Fauzi A, Zainal Z, Ismail I. Diverse and dynamic roles
959 of F ‐ box proteins in plant biology. Planta. 2020;251(3):1–31.

960 129. Schumann N, Navarro-quezada A, Ullrich K, Kuhl C, Quint M. Molecular evolution
961 and selection patterns of plant F-box proteins with C-terminal kelch repeats.
962 2011;155(2):835–850.

963 130. Wang S, Zhou H, Wu J, Han J, Li S, Shao S. Transcriptomic analysis reveals genes

964 mediating salt tolerance through calcineurin / CchA-independent signaling in
965 Aspergillus nidulans. Biomed Res Int. 2017.

966 131. Bylund RANO, Persson BC, Lundberg LAC, Wikstro PM. A novel ribosome-
967 associated protein is important for efficient translation in Escherichia coli. J Bacteriol.
968 1997;179(14):4567–74.

969 132. Suzuki S, Tatsuguchi A, Matsumoto E, Kawazoe M, Kaminishi T, Shirouzu M, et al.
970 Structural characterization of the ribosome maturation protein RimM. J Bacteriol.
971 2007;189(17):6397–406.

972 133. Ma L, Li G. FAR1-related sequence (FRS) and FRS-related factor (FRF) family
973 proteins in Arabidopsis growth and development. Front Plant Sci. 2018;9(692).

974 134. Salojärvi J, Smolander O-P, Nieminen K, Rajaraman S, Safronov O, Safdari P, et al.
975 Genome sequencing and population genomic analyses provide insights into the
976 adaptive landscape of silver birch. Nat Genet. 2017;49(6):904–12.

977 135. Han G, Lu C, Guo J, Qiao Z, Sui N, Qiu N. C2H2 zinc finger proteins: master
978 regulators of abiotic stress responses in plants. Front Plant Sci. 2020;11(115).

979 136. K MJ, Singh D, Sharma M, Sharma M, Mannully CT, Shukla BN, et al. The FCS-
980 LIKE ZINC FINGER 6 and 10 are involved in regulating osmotic stress responses in
981 Arabidopsis. Plant Signal Behav. 2019;14(6):1592535.

982 137. Emerson RO, Thomas JH. Adaptive evolution in zinc finger transcription factors. PLoS
983 Genet. 2009;5(1):e1000325.

984 138. Mukherjee K, Brocchieri L. Evolution of Plant Homeobox Genes. eLS. 2010;

985 139. Khan N, Hu C, Khan WA, Wang W, Ke H, Huijie D, et al. Genome-wide
986 identification, classification, and expression pattern of homeobox gene family in

Brassica rapa under various stresses. Sci Rep. 2018;8(1):1–17.

140. Trisolini L, Gambacorta N, Gorgoglione R, Montaruli M, Laera L, Colella F, et al. FAD/NADH dependent oxidoreductases: from different amino acid sequences to similar protein shapes for playing an ancient function. J Clin Med. 2019;8(12):2117.

141. Cao Y, Ma L. To splice or to transcribe: SKIP-mediated environmental fitness and development in plants. Front Plant Sci. 2019;10:1222.

Supporting Information file

Table S1.xls

List of species and location of their collection

Table S2.xls

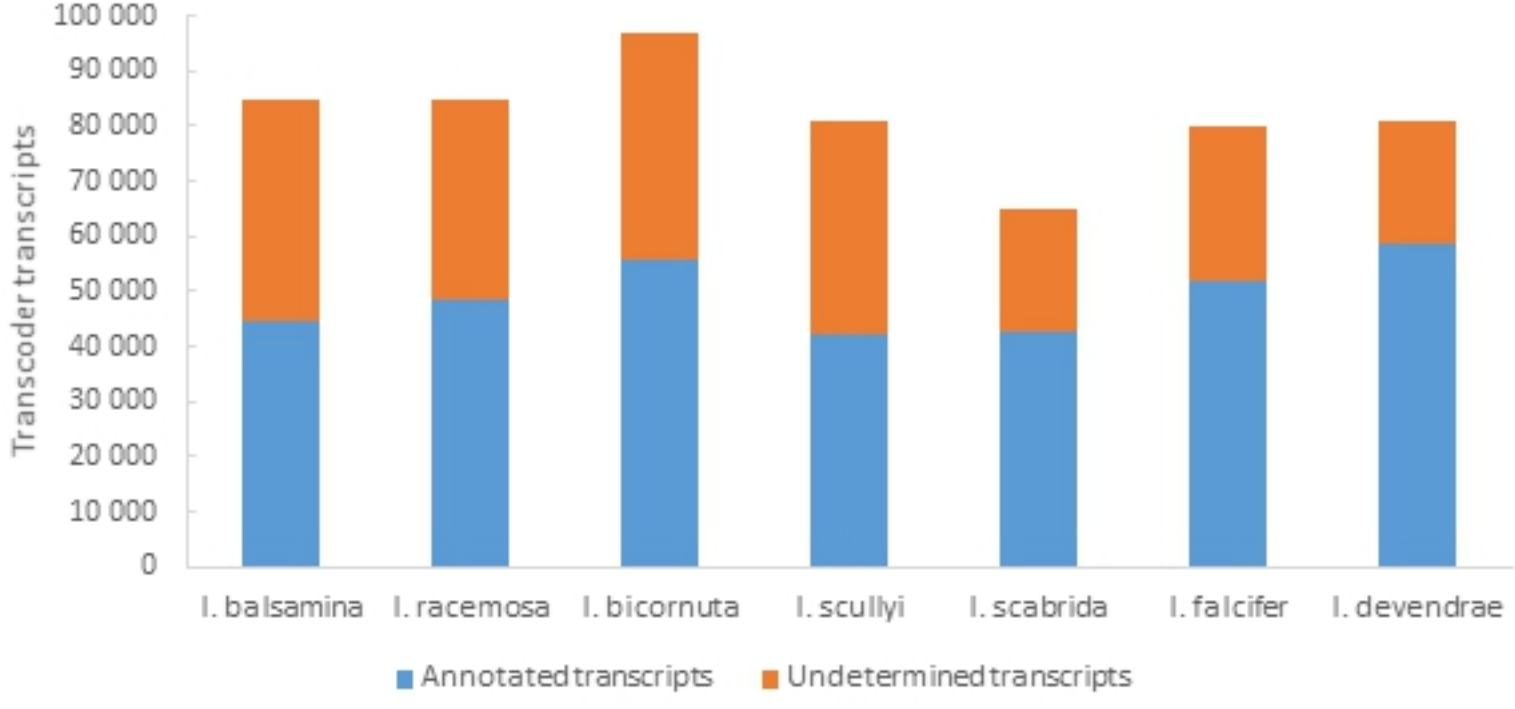
Identified GO terms for each species.

Table S3.xls

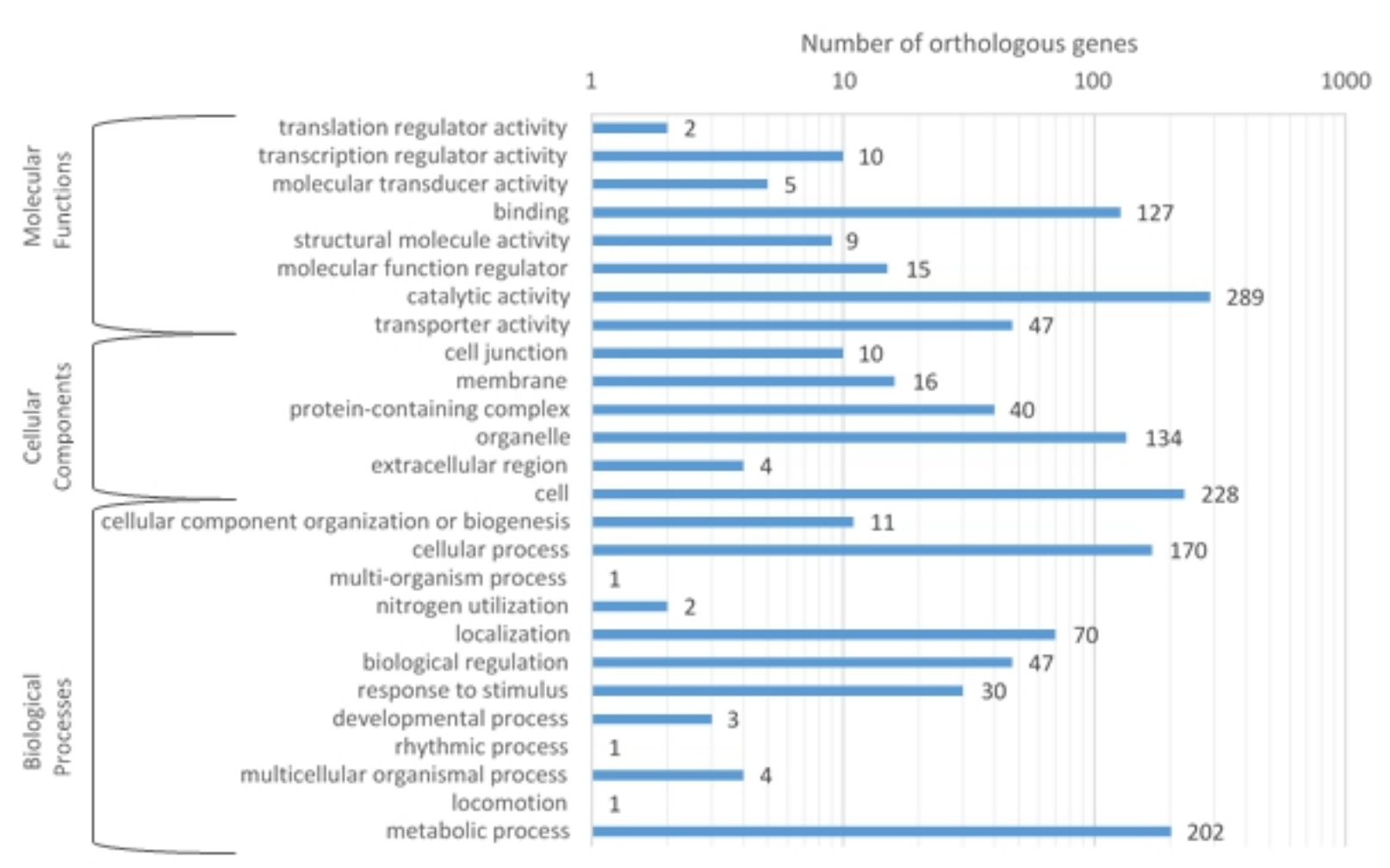
Identified orthologues under selection for each species.

Table S4.xls

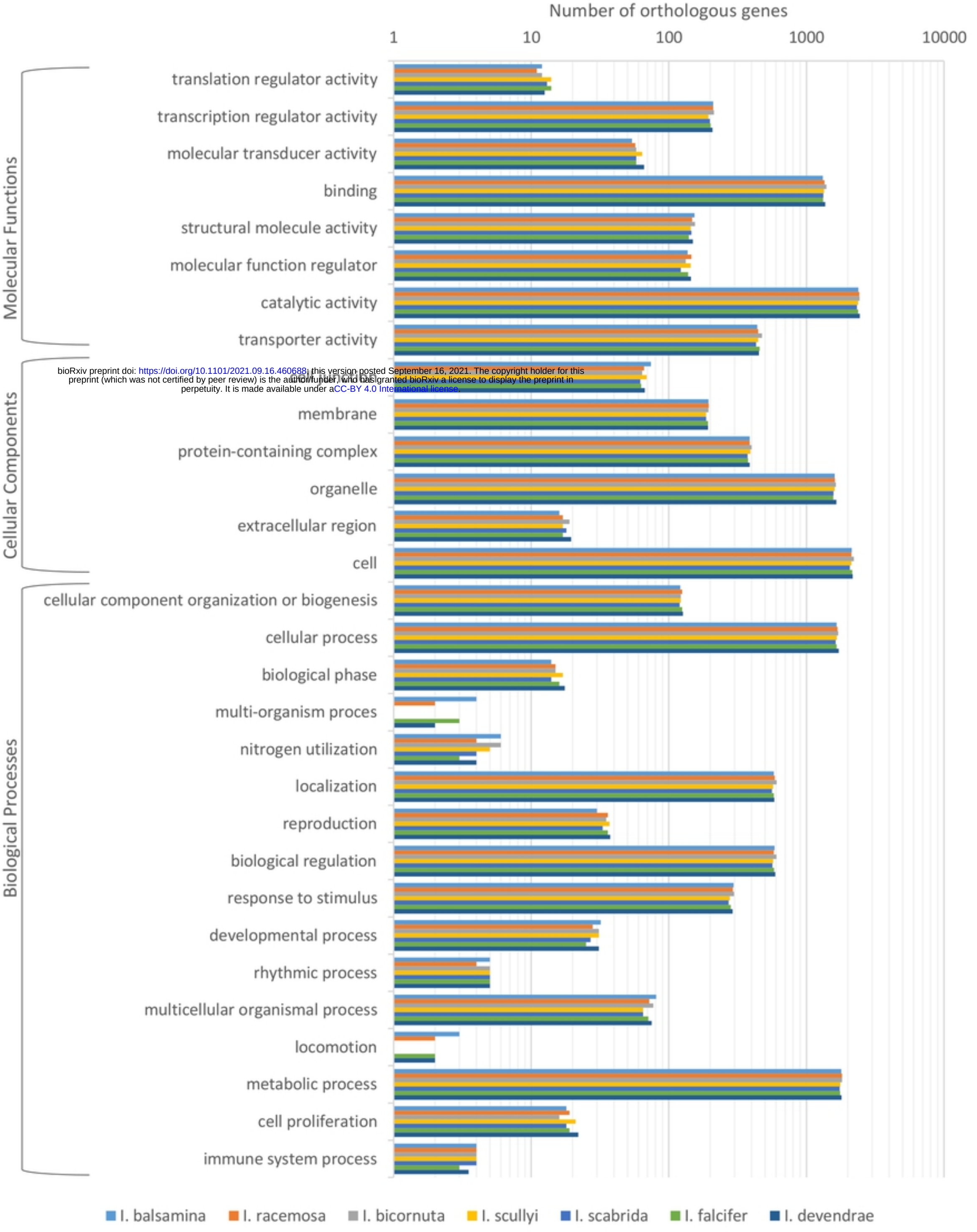
Identified orthologues under selection common for all species.


100 000
90 000
80 000
70 000
60 000
50 000
40 000
30 000
20 000
10 000
0
Transcoder transcripts
I. balsamina
I. racemosa
I. bicornuta
I. scullyi
I. scabrida
I. falcifer
I. devendrae
Annotated transcripts
Undetermined transcripts

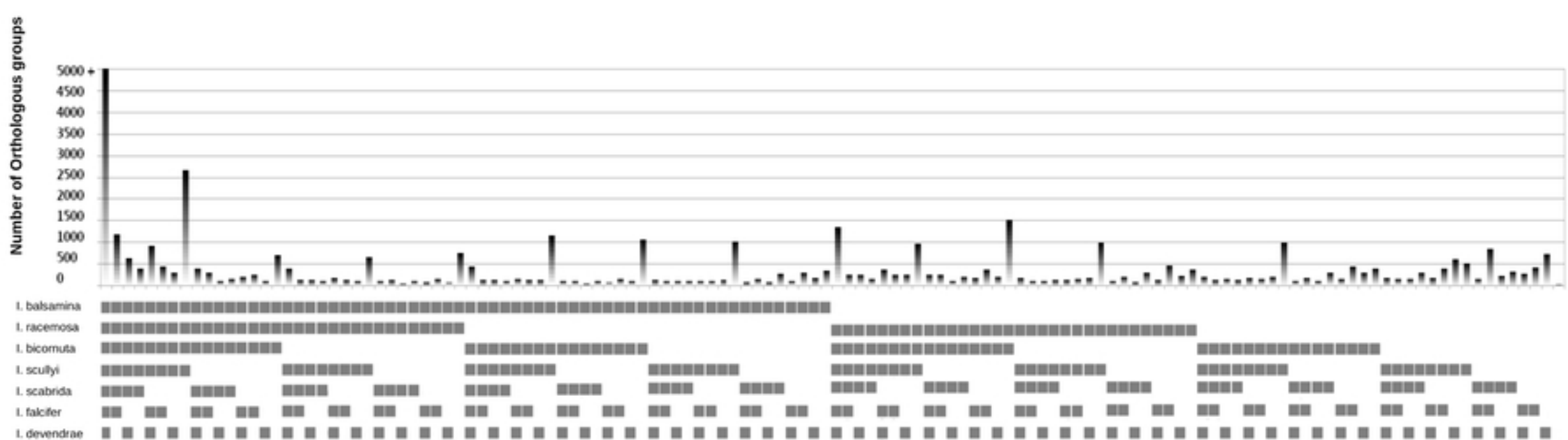
Figure


Number of orthologous genes
1
10
100
1000
Molecular Functions
Cellular Components
Biological Processes
translation regulator activity
transcription regulator activity
molecular transducer activity
binding
structural molecule activity
molecular function regulator
catalytic activity
transporter activity
cell junction
membrane
protein-containing complex
organelle
extracellular region
cell
cellular component organization or biogenesis
cellular process
multi-organism process
nitrogen utilization
localization
biological regulation
response to stimulus
developmental process
rhythmic process
multicellular organismal process
locomotion
metabolic process
2
10
5
127
9
15
289
47
10
16
40
134
4
228
11
170
1
2
70
47
30
3
1
4
1
202

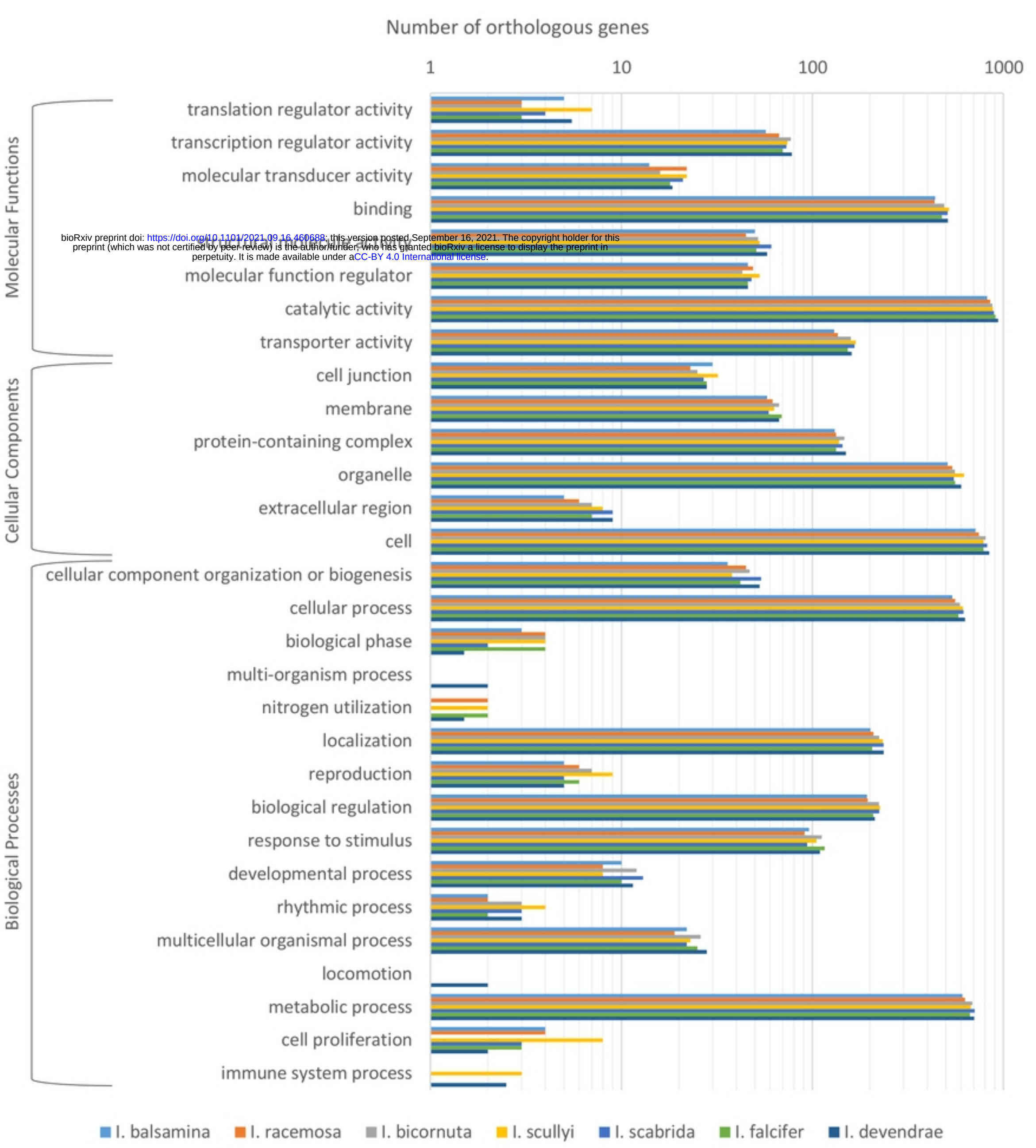
Figure



Figure


Number of Orthologous groups
5000 +
4500
4000
3500
3000
2500
2000
1500
1000
500
0
I. balsamina
I. racemosa
I. bicornuta
I. scullyi
I. scabrida
I. falcifer
I. devendrae

Figure



Figure