High-resolution transcriptome atlas and improved genome assembly of common buckwheat, *Fagopyrum esculentum*.

Aleksey A. Penin[1], Artem S. Kasianov[1], Anna V. Klepikova[1], Ilya V. Kirov[2], Evgeny S. Gerasimov[3], Aleksey N. Fesenko[4], Maria D. Logacheva[1,5]*

[1] - Institute for Information Transmission Problems of the Russian Academy of Sciences, 127051 Moscow, Russia
[2] – All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia
[3] – Lomonosov Moscow State University, Biological faculty, Moscow, Russia
[4] - Federal Scientific Center of Legumes and Groat Crops, 302502 Orel, Russia
[5] - Skolkovo Institute of Science and Technology, 143026 Moscow, Russia
* corresponding author, maria.log@gmail.com

Keywords: *Fagopyrum esculentum*, buckwheat, genome, transcriptome map, Caryophyllales, petal, transposable elements.
Running title: Transcriptome map and genome assembly of common buckwheat

Summary.

Common buckwheat (*Fagopyrum esculentum*) is an important non-cereal grain crop and a prospective component of functional food. Despite this, the genomic resources for this species and for the whole family Polygonaceae, to which it belongs, are scarce. Here, we report the assembly of the buckwheat genome using long-read technology and a high-resolution expression atlas including 46 organs and developmental stages. We found that the buckwheat genome has an extremely high content of transposable elements, including several classes of recently (0.5-1 Mya) multiplied TEs ("transposon burst") and gradually accumulated TEs. The difference in TE content is a major factor contributing to the 3-fold increase in the genome size of *F. esculentum* compared with its sister species *F. tataricum*. Moreover, we detected the differences in TE content between the wild ancestral subspecies *F. esculentum* ssp. *ancestrale* and buckwheat cultivars, suggesting that TE activity accompanied buckwheat domestication. Expression profiling allowed us to test a hypothesis about the genetic control of petaloidy in buckwheat. We showed that it is not mediated by B-class gene activity, in contrast to the prediction from the ABC model. Based on a survey of expression profiles and phylogenetic analysis, we identified the MYB family transcription factor gene tr_18111 as a potential candidate for the determination of conical cells in buckwheat petaloid tepals. The information on expression patterns has been integrated into the publicly available database TraVA: http://travadb.org/browse/Species=Fesc/. The improved genome assembly and transcriptomic resources will enable research on buckwheat, including practical applications.

Introduction.

Common buckwheat (*Fagopyrum esculentum*) is an important non-cereal grain crop of particular importance in Russia, China and Ukraine, with production of ~ 2 million tons. With the growing interest in healthy lifestyles, buckwheat has received a great deal of attention worldwide as a component of functional food in recent years (for review, see (Giménez-Bastida and Zieliński, 2015; Alvarez-Jubete *et al.*, 2010)). Buckwheat is highly resistant to soil contamination by aluminium and lead and is a promising phytoremediator (Ma *et al.*, 1997; Honda *et al.*, 2007). Its productivity as an agricultural plant is, however, limited by several unfavourable traits: obligate insect outcrossing pollination, an extended period of flowering and susceptibility to drought and freezing. Novel methods of breeding that involve marker-assisted and genomic selection have the potential to overcome these limitations. For this purpose, genomic information is highly desired. A draft assembly of the buckwheat genome was reported in 2016 (Yasui *et al.*, 2016); however, it is based only on Illumina technology and shows moderate continuity (contig N50 1 Kb, scaffold N50 = 25 Kb). This assembly is likely to encompass most coding regions and is suitable for the analysis of single genes (though with some limitations, see, e.g., Lei *et al.*, 2017). Many researchers opt not to use this assembly as a reference for transcriptomic studies on buckwheat and instead generate their own de novo transcriptome assemblies (Xu *et al.*, 2017; X., Fang *et al.*, 2019; Z., Fang *et al.*, 2019). This situation decreases the accuracy of gene expression estimates and hampers the comparison and meta-analysis of the results. Thus, the generation of a more contiguous assembly for the buckwheat genome complemented with transcriptomic information is urgently needed (Joshi *et al.*, 2019).

Sequencing technologies have been responsible for tremendous progress; one of the key innovations of these technologies in recent years was the development of NGS platforms capable of generating long reads. This is of special importance for plant scientists because plant genomes are large and complex and are shaped by multiple segmental and whole-genome duplications and transposable element (TE) activity. Thus, they cannot be reliably assembled using 2nd-generation NGS technologies, which generate reads with a maximum length of 100-300 bp. The utility of long-read technology for the assembly of complex genomes has been demonstrated in several plant species (Schmidt *et al.*, 2017; Murigneux *et al.*, 2020). Long-read technologies have encouraged scientists to conduct studies that were impossible even with the gold-standard quality genome of *Arabidopsis thaliana* due to the complex rearrangements caused by transposition. On this basis, we performed the sequencing, de novo assembly and annotation of the buckwheat genome using 3rd-generation technologies – SMRT and nanopore sequencing – and generated a transcriptome map of *F. esculentum* from 46 organs and developmental stages. We demonstrate the utility of detailed transcriptome profiling for testing and generating hypotheses about the involvement of genes in specific processes, particularly the genetic basis of floral organ identity and anthocyanin synthesis. We show that gene expression patterns are a valuable addition to sequence-based information.

We expect that this newly generated genome and transcriptome resource will be useful for the development of buckwheat genetics and breeding. Buckwheat belongs to an isolated group of flowering plants, the non-core Caryophyllales, which are distantly related to model species (Yao *et al.*, 2019). While their sister clade, the core Caryophyllales, contains several plants with well-characterized genomes (sugarbeet (Dohm *et al.*, 2013) and quinoa (Jarvis *et al.*, 2017)), genomic and transcriptomic resources for the non-core Caryophyllales are lacking. Thus, the development of a high-resolution transcriptome map will provide a resource for comparative transcriptomic analyses in the phylogenetic context.

Results and discussion.

Assembly and annotation.

The Dasha cultivar was selected for reference genome assembly. Dasha is a recently developed fast-growing determinate cultivar characterized by resistance to lodging and a high photosynthesis efficiency (Fesenko *et al.*, 2018). The assembly process consisted of three stages: contig assembly, scaffolding and gap closing. For contig assembly, a Newbler assembler was used. This assembler, initially developed for 454 data, also performs well for other types of sequencing data, including merged Illumina paired 250-bp or 300-bp reads. The 250-bp Illumina reads and mate-pair reads provided a backbone for the assembly; further improvement of the assembly, including gap closing, was performed using long reads generated by SMRT CCS technology on the Pacific Bioscience platform (Table S1). The final assembly consisted of 88 078 scaffolds with a total length of 1.2 Gb and N50 ~ 180 Kb (Table S2). The N50 of this assembly represents a 39x improvement compared with that of the previous assembly at the level of contigs. A search of the conserved gene set (BUSCOs) also showed high continuity of the assembly: the percentage of complete genes was 98.6%, and those of fragmented and missing genes were both 0.7%. For comparison, these metrics for the previous assembly were 84.2%, 11.5 and 4.3% (Table S3). The annotation yielded 29 514 genes; the annotation metrics are summarized in Table S4. The majority of the genes (97.5%) were supported by RNA-seq (Table S4 and text below). The parameters of the coding fraction of the genes (CDS length, exon length) of buckwheat were similar to those of other plants, while the introns of buckwheat were longer, which is in accord with its larger genome size (Figure 1a). The number of genes was close to that found in sugar beet (Dohm *et al.*, 2013), which belongs to Caryophyllales along with buckwheat, and in *A. thaliana* and was generally within the range typical for the genomes of plants that have not undergone a recent whole-genome duplication. A total of 24 765 buckwheat genes were classified into 12 627 orthogroups. Notably, the buckwheat genome contains a smaller number of orthogroups that include one gene and a larger number of orthogroups that include two or more genes (Figure 1b). This pattern is not, however, compatible with that observed in tetraploids (see, e.g., Leushkin *et al.*, 2013); taking into account the unusually high content and activity of TEs in the buckwheat genome (see below), we conclude that it is a result of segmental duplications caused by TEs. The distribution of gene ontology categories was similar between buckwheat and *Arabidopsis* (except for the “secondary metabolic process” category, which includes more genes in *Arabidopsis*, presumably due to its ability to synthesize glucosinolates) (Figure 1c). Additionally, a search for GO enrichment in the genes belonging to the 1-to-many orthogroups (one gene in *Arabidopsis* and two or more in buckwheat) did not reveal any enrichment (FDR < 0.01). This shows that duplications are not confined to specific gene groups. A total of 1 425 buckwheat genes were classified as transcription factors (TFs, Table S5); the total number and distribution of TF classes were similar to those of other caryophyllids (1 058 in *Beta vulgaris*, 1 259 in *Amaranthus hypochondriacus*) and *A. thaliana* (1 717). The extranuclear genomes – plastid and mitochondrial – were assembled using different procedures and are reported elsewhere (Logacheva *et al.*, 2020).

Transposable elements and the evolution of genome size in the genus *Fagopyrum*.

The genus *Fagopyrum* includes approximately 25 species, which fall into two clades: cymosum (large achene) and urophyllum (small achene) (reviewed in (Ohsako and Li, 2020)). The genome sizes of the two clades vary significantly despite showing the same chromosome number in most species (Nagano *et al.*, 2000); the basis of this variation is unknown. *F. esculentum* is a member of the large-seeded clade; its genome is ~3 times larger (1.39 pg/1C) than that of the closely related species *F. tataricum* (0.56 pg/1C), while the chromosome number is the same in these species (Neethirajan *et al.*, 2011). Polyploidy and TEs are the main drivers of plant genome size variability across species (Bennett and Leitch, 2005; Vitte and Panaud, 2005). Because both species are diploids and they show the same chromosome number (2n=2x=16), we hypothesized that TE activity may be responsible for the increase in *F. esculentum* genome size. To obtain preliminary insight into the genome-wide differences in the repeatome composition between the two species, we performed the comparative clustering of genomic NGS reads followed by cluster annotation using RepeatExplorer software (Novak *et al.*, 2013; Novák *et al.*, 2017). This analysis revealed a similar genome abundance of satellite (3.45% and 4%) and class II (DNA transposons, 1.97% and 1.98%) repeats, while the portion of the genome occupied by retrotransposons (RTEs, class I repeats) was ~2 times greater in *F. esculentum* (64.8%) than in *F. tataricum* (30.7%) (Figure 2a). Then, we took advantage of the current *F. esculentum* genome sequence assembly to obtain deeper insight into LTR RTE diversity and compared these data with those of *F. tataricum* (http://mbkbase.org/Pinku1/FtChromosomeV2.fasta.gz, (Zhang *et al.*, 2017). The de novo prediction of full-length LTR retrotransposons in the two genome assemblies revealed 14 435 and 4 891 sequences for *F. esculentum* and *F. tataricum*, respectively, showing good concordance with the NGS-based estimation of TE abundance differences. The further classification of RTEs indicated that ~70% of all RTEs belonged to the Ty3/Gypsy superfamily. The family-based classification of all RTEs revealed striking differences in the copy numbers of distinct Ty3/Gypsy and Ty1/Copia families between the two species (Figure 2b, c). Athila Ty3/Gypsy was the most prevalent (42%) family in the *F. esculentum* genome, including ~25 times more RTEs (6117) than the number found in *F. tataricum* (241) (Figure 2b). In contrast, 59% of *F. tataricum* RTEs (2911) belong to the Tekay Ty3/gypsy RTEs (Figure 2b), although the number of identified RTEs of this family was almost identical between the two species. Similarly, Ty1/Copia RTEs showed discrepancies in the copy numbers of distinct families between the genomes of two species, with the Ale, SIRE, TAR and Ikeros families being the most prevalent in *F. esculentum*, while the Tork family is the largest family in *F. tataricum,* although the copy number is almost the same between the two genomes. This suggests that the RTEs of several families have accumulated at a higher rate in *F. esculentum* than in *F. tataricum*. We then asked whether the insertion time of RTE activity differed between these species. The obtained estimates clearly showed significant differences (Fisher's exact test for count data, p-value < 2.2e-16) between the species, with 26.6% (3 834 complete RTEs) of the RTEs of the *F. esculentum* genome having been inserted during the last 0.5 million years ('recent' insertions), while only 9.6% of *F. tataricum* RTE insertions were classified as 'recent' insertions. Surprisingly, the comparison of insertion times for different RTE families revealed that Athila accumulation was characterized by gradual dynamics in both species, without a burst of recent activity in *F. esculentum* (Figure 2e). A similar situation was observed for the CRM and TAR families. In contrast, the Ale, Ikeros, and Tork families and the SIRE family showed signatures of bursts of recent activity in *F. esculentum* and *F. tataricum*, respectively. Thus, the greater *F. esculentum* genome size compared with that of *F. tataricum* is influenced by the generally higher rate of transposition activity and gradual accumulation of Athila members, together with the burst of recent activity of Ale, Ikeros and Tork family members. Along with the comparison of *F. tataricum* and *F. esculentum*, which allowed us to reveal the basis of the rapid genome size change, we surveyed the TE content within *F. esculentum*. We analysed WGS data for 10 cultivars and the ancestral subspecies *F. esculentum* ssp. *ancestrale* and found that the abundance of TEs differed between the cultivars and *F. esculentum* ssp. *ancestrale* and between the cultivars. (Figure S1). These differences were correlated with the breeding history of the cultivars; for example, the Kazanka (Kaz) and Dialog (Dia) cultivars, which were clustered together in the TE abundance tree (Figure S2), are indeed closely related, with Kazanka being one of the progenitors of Dialog (Fesenko, 2009).

Buckwheat transcriptome atlas and FescTraVA database.

Gene expression information is integral for many areas of plant biology; in particular, it helps to reveal the role of genes in growth, development and stress response (see e.g. (Zhao *et al.*, 2014), (Shumayla *et al.*, 2017)), to find the loci associated with agriculturally important traits (Galpaz *et al.*, 2018), to infer phylogenetic relationships between species (Guo *et al.*, 2020). With this premise we constructed transcriptome map of buckwheat based on newly generated genome sequence and annotation. We selected 46 samples from different organs and developmental stages of buckwheat, ranging from seeds and seedlings to senescent organs (Table S6). Since buckwheat exhibits two forms of flowers – pin (long style and short stamens) and thrum (short style and long stamens) – most samples of flowers and floral parts were collected separately for the two forms. The transcriptome is highly dynamic, especially in plants, in which it depends greatly on environmental conditions and the circadian cycle. Taking this into account, the growing and collection of samples for the transcriptome map was performed under the same conditions and at the same time of the day. For each sample, > 20 million reads were obtained, and the average number of reads per sample was 12 million. The correlation coefficient of the samples lies in the range 0.85-1.00 with an average of 0.96 (Table S7). The clustering of the samples reflects their biological similarity (Figure 3a). The number of genes expressed in each sample was similar, and the greatest numbers were observed in seeds and flowers (Figure 3b). The number of genes expressed in at least one sample was 28 770 (97% of annotated genes), while 12 172 genes were expressed in all samples, and 744 genes were not expressed in any (Figure 3c). The general patterns of gene expression were similar to those observed in *Arabidopsis* (Klepikova *et al.*, 2016). The estimation of gene expression breadth using Shannon entropy showed that most genes exhibited broad expression patterns (Figure S2). The genes in the fraction with the broadest expression patterns were enriched in the GO categories “photoperiodism, flowering” and “protein transporter activity”, while those in the fraction with the narrowest expression patterns were enriched in “serine-type endopeptidase inhibitor activity”, “negative regulation of endopeptidase activity”, and “endomembrane system” (Table S8).

RNA-seq data, especially high-resolution transcriptome maps, are an important resource for the identification of stably expressed genes, which are useful as reference genes for qRT-PCR (Zhou *et al.*, 2017; Machado *et al.*, 2020). We used the standard deviation of gene expression divided by mean expression (standard deviation; SD)/mean) as a measure of expression stability, and we identified 20 transcripts with SD/mean ratios < 0.2, 205 with ratios < 0.25, and 720 with ratios < 0.3. In a previous study, several gene orthologues of stably expressed *Arabidopsis* genes (Czechowski *et al.*, 2005) were evaluated in five organs of buckwheat using qRT-PCR (Demidenko *et al.*, 2011). Ten transcripts corresponded to these genes, and only one of them – tr_15881 (an orthologue of AT4G34270) – was stably expressed. The organs where expression showed the greatest differences were the anthers and roots (Figure S3), which were not sampled in the abovementioned study. This stresses the importance of balanced and high-resolution transcriptome maps for the assessment of gene stability.

To make the information on gene expression levels readily available to the research community, we summarized the results in the TraVA database: http://travadb.org/browse/Species=Fesc/. This database uses the same intuitive user-friendly interface as the transcriptome atlas of *A. thaliana* published earlier (Klepikova *et al.*, 2016). For convenience, the database can be searched according to either the identifier of the buckwheat transcript or the identifier (common name or AT*G* TAIR identifier) of the homologous *A. thaliana* gene. When the *A. thaliana* gene is a member of an orthogroup including several buckwheat genes, the expression profiles for all of these genes are shown (Figure 3d, e). The data can be downloaded in PNG and XLS formats.

The potential of comparative transcriptomics for the inference of gene function.

High-resolution gene expression maps are a useful tool for the generation and testing of hypotheses about gene function in non-model organisms. A commonly used approach is the identification of an orthologue of a gene of interest in a model organism, where the functions of the orthologues are assumed to be similar (the so-called orthologue conjecture) (Gabaldón and Koonin, 2013; Altenhoff *et al.*, 2012). Thus, the efforts of comparative genomics are focused on the identification of orthologues (Tulpan and Leger, 2017). However, this approach has several limitations. The most critical aspect of plant science is the complexity of plant gene families, which are shaped by multiple whole-genome and segmental duplications. This leads to the impossibility of inferring 1-to-1 orthologues and makes it necessary to work with orthogroups – the sets of genes descended from a single gene in the last common ancestor. The diversification of gene function within an orthogroup is a common way in which new structures and pathways emerge in plants (e.g., Yamaguchi *et al.*, 2006; Ober, 2005). However, based on sequence similarity, it is usually impossible to identify genes that share the same function in two species because subtle differences at the coding sequence level can contribute greatly to changes in function (a single amino acid change can turn a repressor to an activator) (Hanzawa *et al.*, 2005).

The transcriptome atlas enables the comparison of the expression profiles of the genes within an orthogroup, thus allowing the prioritization of the hypotheses regarding their functional correspondence. In a previous study, we constructed transcriptome atlases for *Arabidopsis* (Klepikova *et al.*, 2016) and tomato (Penin *et al.*, 2019) using the same approaches and a similar set of samples employed for buckwheat; these species represent two major groups of eudicots (rosids and asterids) and provide a basis for the comparative analysis of gene expression profiles. As an illustration of the novel information that the atlas can provide, we will focus on at the orthogroup including the *APETALA1* gene. The buckwheat genome carries three orthologues of the *AP1* gene. They share high similarity (74-77% at the nucleotide sequence level) and likely arose via whole-genome or segmental duplication after the diversification of core and non-core Caryophyllales (Figure 4a). The comparison of their expression profiles showed that the expression of all of these genes was confined to the flower; however, the patterns and levels of expression were different (Figure 4a). tr_22431 showed the highest expression level in tepals (the structures surrounding the reproductive organs of the flower in buckwheat, which are similar to petals and sepals; see discussion below), while two others showed the highest expression in the anthers and the receptacle of the developing flower. In *Arabidopsis*, *AP1* is known to regulate sepal and petal development (Irish and Sussex, 1990, p.1). Based on the expression profiles, the most plausible candidate for this function was tr_22431. Indeed, a study by Liu *et al.* (2019), who performed a complementation assay with an *Arabidopsis ap1* mutant and a buckwheat *AP1* orthologue (corresponding to tr_8095 from our annotation, 100% identity), showed a limited conservation of function between *Arabidopsis AP1* and this buckwheat gene. This shows the congruence of functional studies and expression profiles and calls for the widespread application of the buckwheat transcriptome atlas for functional genomics.

Another example is provided by the buckwheat tr_14963 gene, which encodes an ~2000 aa protein with high similarity to the DNA polymerase epsilon subunit. It exists in a single copy in most plants, excluding those that recently underwent a whole-genome duplication (Figure 4b). However, in *A. thaliana,* there are two genes, *TIL1* (*AT2G27120*) and *TIL2* (*AT2G27120*), corresponding to this single gene. They show very close similarity values (56%) to the buckwheat orthologue, making it impossible to determine which of the *TIL* genes corresponds functionally to the buckwheat tr_14963 gene. In contrast, the expression profiles show clear differentiation: *TIL1* presents a more similar pattern to the buckwheat gene, where the maximal expression level is associated with growing tissues, while *TIL2* exhibits a narrow pattern and is expressed almost exclusively in the anthers (Figure 4b).

The transcriptome atlas provides a clue regarding the origins of petaloidy in buckwheat.

An important characteristic of flowering plants is the perianth – the organs that surround the reproductive structures of the flower and functions in protection and/or the attraction of pollinators. The perianth can be either double, being differentiated into a calyx (sepals, usually small and green, protect the developing floral organs) and corolla (petals, large and showy, attract pollinators) or undifferentiated, in which a single type of organ resembles either petals or sepals. A double perianth is typical for eudicots (including the model species *A. thaliana*), while an undifferentiated perianth is usually found in monocots. The well-known ABC model of floral organ identity suggests that this structure is controlled by the combinatorial action of three classes of TFs – A, B and C. According to this model, sepal identity is mediated by the A class genes, carpel identity by C class, petal identity by classes A and B and stamen identity by classes B and C (Coen and Meyerowitz, 1991). The undifferentiated petaloid perianth has been hypothesized to be attributable to the expansion of the B class (van Tunen, Arjen *et al.*, 1993); this hypothesis has been experimentally corroborated in several monocot species (Kanno *et al.*, 2003; Nakamura *et al.*, 2005). In Caryophyllales, perianth structure is highly diverse; this group includes species with both double and undifferentiated perianths; the latter type is presumably an ancestral character for the whole order (Brockington *et al.*, 2009). Buckwheat and other species from the family Polygonaceae exhibit an undifferentiated perianth. In *F. esculentum,* the perianth is petal like, white or pink, while in several other buckwheat species, it is green, sepal like. A plausible hypothesis is that in *F. esculentum,* class A and B genes are active in the perianth, while in species with a sepal-like perianth, only class A genes are active. The availability of the genome sequence in combination with a detailed transcriptome map allows this hypothesis to be tested. The search for orthologues of B-class genes showed that the buckwheat genome carries four orthologues of *AP3* and one orthologue of *PI* (Figure 5). We searched for their expression in different floral organs, and we found that B is not expressed or is expressed at background levels in the perianth or only in stamens. In tomato, the expression pattern of B-class genes is compatible with the ABC model; the conservation of the function of *AP3* and *PI* orthologues (the latter is duplicated in tomato) was previously shown in direct RNAi experiments (Guo *et al.*, 2016; de Martino *et al.*, 2006). Tomato belongs to the asterids, while *Arabidopsis* belongs to the rosids. Buckwheat is more closely related to the asterids, which suggests that the control of petal identity by the B-class genes was lost in an ancestor of buckwheat after its divergence from asterids. Earlier, we suggest that B-class genes are not involved in petaloidy in buckwheat based on the phenotype of a buckwheat homeotic mutant (Logacheva *et al.*, 2008); this hypothesis now gains support from the obtained expression profiles. Studies on the complementation of *Arabidopsis ap3* and *pi* mutants via the overexpression of buckwheat *AP3* and *PI* genes also show limited conservation of function (Fang *et al.*, 2015; Fang *et al.*, 2014). An alternative programme of petal development, which is not dependent on B-class genes, was also found in a family of the core Caryophyllales, Aizoaceae (Brockington *et al.*, 2012).

The most notable features of petaloidy in buckwheat tepals are white colour (while the sepals are usually green) and the shape of epidermal cells, which are conical (Hong *et al.*, 1998), as found in the petals of most plants (Whitney *et al.*, 2011). Studies in the model organism *Antirrhinum majus* have shown that the major role in the development of conical cells is played by the *MIXTA* gene, encoding an R2R3-MYB transcription factor (Glover *et al.*, 1998; Perez-Rodriguez, 2005). Additional factors from the MYB family that are closely related to *MIXTA* control other aspects of this process (Perez-Rodriguez, 2005). *Arabidopsis* does not have a *MIXTA* orthologue, but two genes from a sister clade, *MYB106* (*AT3G01140*) and *MYB16* (*AT5G15310*), as well as several other MYB family TFs, participate in the control of epidermal cell development (Oshima *et al.*, 2013). The evolutionary history of *MIXTA*/*MIXTA-like* genes is complicated, being shaped by multiple duplications and gene losses, particularly in the *MIXTA* clade (Brockington *et al.*, 2013). In the buckwheat genome, we identified 118 genes with significant similarity to MIXTA/MYB16/MYB106 (Table S9). Six of them showed an expression pattern compatible with their role in the determination of conical cells, with prevalent expression in developing tepals (Figure S4). One of these transcripts – tr_18111 – exhibit a protein motif that is characteristic of MIXTA, MYB16 and MYB106. Phylogenetic analysis also indicates that tr_18111 falls into the clade of *MIXTA-like* genes (Figure S5). This strongly supports the hypothesis that petal conical cell identity is controlled by *MIXTA/MIXTA-LIKE* genes in buckwheat as well. Within *Antirrhinum, MIXTA* has been shown to be under the control of B-class genes in *A. majus* (Perez-Rodriguez, 2005; Manchado-Rojo *et al.*, 2012), and this regulatory contour seems to be conserved as far as the monocots (Pan *et al.*, 2014). The absence of B-class gene activity in buckwheat tepals along with the presence of conical cells suggests that this link is not universal. Notably, the morphology of tepal epidermal cells differs within the genus *Fagopyrum*: they are conical in *F. esculentum* and *F. tataricum* (though the latter has green “sepaloid” tepals), whereas they are elongated with sinuated walls in other species (Hong *et al.*, 1998). This genus is thus a good model for the study of the genetic control of petal identity and petal epidermal cell structure. The characterization of the genome and expression profiles of *F. esculentum* enables a comparative transcriptomic study of the genus.

Conclusions.

We characterized the 1.5 Gb genome of *F. esculentum*, providing a reference assembly with high contiguity and an improved representation of protein-coding genes. The genome size of *F. esculentum*, which is three times larger than that of its sister species *F. tataricum*, is a product of a “transposon burst” that occurred 0.5-1 Mya. To provide a framework for the functional and comparative genomics of buckwheat, we constructed a comprehensive transcriptome atlas from 46 tissues, organs and developmental stages. We demonstrated that high-resolution transcriptome maps provide new information that allows us to discern the function of closely related genes and to test biological hypotheses regarding gene function and the conservation of developmental programmes.

Experimental procedures.

Genome sequencing, assembly and assembly quality check.

The Dasha cultivar was chosen as the source material for the construction of the reference genome. DNA was extracted from fresh-frozen young leaves free of visible injuries, necrosis and/or presence of pathogens using the CTAB protocol (Doyle and Doyle, 1987). For Illumina sequencing, a fragment library was prepared using the TruSeq DNA sample preparation kit (Illumina), with 500 ng of DNA as the input and size selection of fragments in the range of 500–600 bp (corresponding to a 380-480 bp insert length) in 2% agarose gels. Sequencing was performed on a HiSeq2500 instrument with HiSeq Rapid 500 cycling reagents at the University of Illinois at Urbana-Champaign (Roy J. Carver Biotechnology Center). Mate-pair libraries were prepared with a Nextera Mate-pair sample preparation kit (Illumina) with 5000 ng of DNA as the input. Size selection was performed in agarose gels; three lengths were selected: 3–4 kbp, 5–7 kbp and 8–10 kbp. Mate-pair libraries were sequenced on a HiSeq2000 instrument with a TruSeq SBS 200 cycle kit. For sequencing on the Pacific Bioscience platform, DNA was extracted in the same way indicated above and sequenced with a SMRT cell on a Sequel II instrument with CCS read settings at the DNALink facility.

The assembly process consisted of three stages: contig assembly, scaffolding and gap closing. For contig assembly, a Newbler assembler was used. Newbler v. 2.9 was run on a 4 Intel(R) Xeon(R) CPU E7-4830 v2 2.20 GHz computer with 2 TB RAM with the following parameters: "-large -ml 150 -mi 85 -het". These parameters specified a minimal overlap length of not less than 150 bp and minimal overlap identity of 85%. The "long" parameter indicates a large genome (>100 Mb), and the "het" parameter indicates heterozygosity. The resulting contigs with lengths greater than 1000 bp were retained for scaffolding. Scaffolding was performed using Platanus v. 1.0.0 (Kajitani *et al.*, 2014) (subprogramme scaffold) with the parameter "-l 3", indicating that contig joining should be supported by at least 3 mate-pair links. The closing of the gaps was carried out in two stages. In the first stage, gaps were closed with Platanus version 1.0.0 software (with default parameters) and Illumina mate-pair data. In the second stage, gaps were closed with PacBio CCS data and LR_Gapcloser software (Xu *et al.*, 2019) with the parameter "-r 5".

RNA extraction and sequencing.

For RNA-seq analysis, RNA was extracted from a set of different organs and developmental stages using the RNEasy Plant Mini kit (Qiagen, Netherlands) with the modifications described previously (Logacheva *et al.*, 2011). RNA quality was checked using a Bioanalyzer2100 (Agilent, USA); samples with RIN > 7 were employed for subsequent analysis. For Oxford Nanopore Technologies sequencing, RNA was converted to cDNA using a Mint cDNA synthesis kit (Evrogen, Russia). The primers for reverse transcription included custom barcodes unique to each sample; thus, the samples corresponding to different organs/stages were pooled after reverse transcription. cDNA was amplified via 24 cycles of amplification. Amplified cDNA was used as the input for library preparation via a standard protocol for genomic DNA with an LSK-309 kit. Sequencing was performed on a MinION system with a 9.5.1 flow cell, and base calling was performed with Guppy. For Illumina sequencing, the libraries were prepared using a TruSeq RNA sample preparation kit (Illumina, USA) following the manufacturer's protocol and sequenced on a HiSeq2000 instrument (Illumina, USA) with 50-bp single reads.

Annotation.

The assembled scaffolds after the first stage of the gap-closing procedure for *F. esculentum* were repeat masked with *RepeatMasker* (ver. 3.3.0) before annotation (http://www.repeatmasker.org). Gene models that overlapped with any of the repeat regions were given a score penalty. RNA-seq data were mapped with *STAR* (ver. 2.4) on the masked genome assembly, and the obtained alignment files were used to generate hints for *Augustus* consisting of hints about the intron start and end sites and transcribed regions. The initial set of genes was predicted ab initio by running the *Augustus* (ver. 3.0.1) gene prediction program with different parameters (species model: *Arabidopsis*/tomato; hints: with/without hints, 4 runs in total) and *GeneMark*. ES (ver. 2.3). All predicted gene models were combined. All predicted gene models were evaluated and ranked according to the following criteria: 1) a small positive bonus was added to the gene model if the intron or exon was supported by RNA-seq data; 2) a positive bonus proportional to the quality and significance of *blastp* alignment with known proteins from the following datasets: *A. thaliana* (maximal bonus if the alignment was significant), SwissProt (smaller bonus) and the NCBI non-redundant database (smallest bonus), was added to the gene model; 3) large negative penalties were applied if the model intersected with repeat regions or showed a hit with a TE. Models were sorted by score. The best-scoring models were picked first, and only one best-scoring isoform was included in the final annotation (no overlapping models). Orthogroups were obtained using

Orthofinder software version 2.4.0 (Emms and Kelly, 2019). The annotation of TFs was performed using PlantRegMap (Tian *et al.*, 2019). The completeness of the assembly was estimated using BUSCO version 4.0.6 software (Simão *et al.*, 2015) for the eudicot lineage and the Viridiplantae lineage.

Transposable element analysis.

To compare the repeatome composition between species, 4 000 000 *F. esculentum* and 1 600 000 *F. tataricum* high-quality paired-end Illumina genomic reads were randomly sampled, and suffixes and prefixes were added using a Python script (https://github.com/Kirovez/RepeatExplorer_scripts/blob/master/prepareReadsREv2.py). The reads were combined into a single FASTA file and used for RepeatExplorer (Novak *et al.*, 2013) and TAREAN analysis (Novák *et al.*, 2017). The locally installed version of RepeatExplorer containing the built-in TAREAN program was run with the following settings: -p -c 150 -C -r 400000000 -P 2. The automatic annotation data were used for repeat classification. Clusters corresponding to organelle DNA were removed, and the portion of the genome consisting of repeats was corrected according to the new number of reads. The table containing the information about cluster annotation and read number was analysed in Rstudio Version 1.2.1335 (http://www.rstudio.com/) with R version 3.6.0 using custom scripts with relevant packages, including ggplot2 (Wickham, 2009) and data.table (https://cran.r-project.org/web/packages/data.table/index.html). For the comparison of the repeatome composition between FE cultivars, 2 000 000 high-quality paired-end Illumina genomic reads were randomly selected, and RepeatExplorer was run as described above. A heatmap was constructed using the ComplexHeatmap R package (Gu *et al.*, 2016).

For the genome-based analysis of LTR RTEs in the two species, the genome assembly of *F. tataricum* (http://mbkbase.org/Pinku1/FtChromosomeV2.fasta.gz, (Zhang *et al.*, 2017)) was downloaded. The LTRharvest tool with default parameters (Ellinghaus *et al.*, 2008) was used to predict putative coordinates of LTR RTEs in the genomes. The obtained gff3 file was sorted and used as the input for the LTRdigest program (Steinbiss *et al.*, 2009) with the following settings: -aaout yes -pptlen 10 30 -pbsoffset 0 3 -pdomevalcutoff 0.001. Hmm repeat profiles of RTE domains were downloaded from the GyDB database (Llorens *et al.*, 2011). The gff3 file from LTRdigest analysis was parsed using a custom python script (https://github.com/Kirovez/LTR-RTE-analysis/blob/master/LtrDiParser_v2.2.py) to extract the sequences of full-length LTR RTEs (those possessing similarity to GAG, RT, RH, AP and INT domains) and their LTRs. TEsorter software (Zhang *et al.*, 2019) was used for RTE classification. The insertion time was calculated with the formula T = k/2r, where k is the distance between LTRs estimated via Kimura's two-parameter method (Kimura, 1980) and r is the mutation rate. A mutation rate of $1.3 \times 10{-}8$ substitutions per site per year (Ma and Bennetzen, 2004) was used. Parameter K was calculated as 0.5 log((1 - 2p -q) * sqrt(1 - 2q)), where p is the transition frequency and q is the transversion frequency. The transition and transversion frequencies were estimated after the alignment of 5' and 3' LTR sequences with clustalw2 software (http://www.clustal.org/clustal2). The insertion time analysis was automated with a custom Python script (https://github.com/Kirovez/LTR-RTE-analysis/blob/master/TEinsertionEstimator.py).

Gene expression analysis.

The total number of reads mapped on a given gene (TGR) was used as a measure of the expression level. The "DESeq" (Anders and Huber, 2010) median approach was used for the normalization of gene expression levels according to library size. We used a threshold of 5 or a more normalized TGR in both biological replicates to identify the expressed genes. "DESeq2" (Love *et al.*, 2014) was used for differential expression analysis with the following settings: a false discovery rate

(FDR) <0.05 and a fold change ≥ 2. The differential expression (DE) score was defined as reported previously (Klepikova *et al.*, 2016). Gene expression pattern width was assessed as described by Penin *et al.* (2019).

Phylogenetic analysis.

Nucleotide sequences were used for phylogenetic analysis. Alignment was performed using MUSCLE (Edgar, 2004), and further was conducted processing using GBlocks with “less stringent” settings (Castresana, 2000). For phylogenetic tree reconstruction, we used IQ-tree (Minh *et al.*, 2020) with the GTR model and 1000 bootstrap replicates.

Data availability statement.

Genome assembly and raw sequencing reads for the Dasha cultivar (reference genome) are available in the NCBI database under Bioproject # PRJNA487881. Raw reads from buckwheat cultivars and *F. esculentum* ssp. *ancestrale* are available under Bioproject # PRJNA627307. The transcriptome reads are available from NCBI SRA, and their accession numbers are listed in Table S5.

Acknowledgments.

The study was supported by Russian Foundation for Basic Research, project # 18-29-13017 (development of a transcriptome map) and Russian Science Foundation, project # 18-76-10018. We are grateful to Dr. Alexey Kondrashov for the access to sequencing and computational facilities.

Author contributions.

AAP constructed transcriptome libraries and coordinated the part of the work on the transcriptome analysis, ASK performed genome assembly, AVK performed gene expression analysis, IVK analysed the repeatome and participated in writing, ESG performed genome annotation, ANF provided plant material, MDL conceived and coordinated the study, constructed genome libraries, performed phylogenetic analysis and drafted the manuscript.

Conflict of interest.

The authors declare that they do not any conflict of interest.

Figures.

Figure 1
a) Annotation of *F. esculentum* compared with those of *A. thaliana* and *S. lycopersicum* in terms of the gene, exon and intron length distribution b) The number of genes per orthogroup in *Fagopyrum esculentum* and *Arabidopsis thaliana* c) The distribution of GO categories (biological function) *in Fagopyrum esculentum* and *Arabidopsis thaliana.*

Figure 2
a) Comparison of genome abundance of different types of repeats in the *F. esculentum* and *F. tataricum* genomes based on the RepeatExplorer (Novak *et al.*, 2013) clustering of NGS genomic reads. b) and c) Number of full-length Ty3/Gypsy and Ty1/Copia LTR RTEs of different families based on prediction in the assembled genome sequences. d) Insertion time comparison for *F. esculentum* (blue) and *F. tataricum* (red) LTR RTEs. e) Cumulative plot of the insertion time for RTE families.

Figure 3.
a) Clustering of the samples (the distance measure is 1 - square of Pearson correlation coefficient, the expression values for replicates are averaged). b) The number of genes expressed in each sample. c) The specificity of gene expression. d) The interface of the FescTraVA database,

showing an example of a search with an *Arabidopsis* TAIR gene ID e) an example of a search with a buckwheat gene ID. Numbers in boxes denote normalized read counts.

Figure 4
a) Phylogeny and expression patterns of the AP1 orthogroup in *Arabidopsis*, buckwheat and tomato. Numbers in boxes denote normalized read counts.
b) Phylogeny and expression patterns of the TIL1-2 orthogroup in *Arabidopsis*, buckwheat and tomato. Numbers in boxes denote normalized read counts.

Figure 5. Phylogeny and expression patterns of the orthologues of the B-class genes *APETALA3* and *PISTILLATA* in *Arabidopsis*, buckwheat and tomato, with a focus on floral organs. Numbers in boxes denote normalized read counts.

Supporting figures.
Figure S1.
Repeat content in *F. esculentum* spp. *ancestrale* and different buckwheat cultivars. DAS – Dasha, BAS – Bashkirskya krasnostebelnaya, DEM – Demetra, DEV – Devyatka, DIA – Dialog, DIZ – Dizajn, KAR – Karadag, KAZ – Kazanka, KUJ – Kujbyshevskaya, DRR/YAS – the sample from the study by Yasui et al. 2016, cultivar unknown, presumably of Japanese origin, KIT/SHI – Shinanonatsusoba, FEA - *F. esculentum* spp. *ancestrale*.

Figure S2
The distribution of Shannon entropy across expressed genes of *F. esculentum.*

Figure S3.
Heatmap showing the expression profiles of the genes – orthologues of the stably expressed genes in *Arabidopsis thaliana*.

Figure S4
Heatmap showing the expression of buckwheat MYB genes sharing significant similarity with MIXTA/MYB16/MYB106.

Figure S5
Phylogenetic tree of angiosperm MIXTA/MIXTA-like genes.

Supporting tables.
Table S1
Data generated for the de novo assembly of the buckwheat genome and accession numbers of raw reads

Table S2
Statistics of the assembly

Table S3
Statistics for the presence of universal single-copy orthologues for the earlier assembly of the buckwheat genome and for the assembly reported in our study

Table S4
Statistics of the protein-coding gene annotation

Table S5
Transcription factor genes annotated in the buckwheat genome

Table S6
List of the samples included in the transcriptome mapping procedure and accession numbers of the raw reads

Table S7
Samples included in the transcriptome mapping procedure and correlations

Table S8
GO enrichment of the genes with broadest and narrowest expression patterns

Table S9
Buckwheat MYB genes selected based on similarity to MIXTA and MYB16/MYB106

References.

Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M. and Dessimoz, C. (2012) Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs J. A. Eisen, ed. *PLoS Computational Biology*, **8**, e1002514.

Alvarez-Jubete, L., Arendt, E.K. and Gallagher, E. (2010) Nutritive value of pseudocereals and their increasing use as functional gluten-free ingredients. *Trends in Food Science & Technology*, **21**, 106–113.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bennett, M.D. and Leitch, I. (2005) Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. *Annals of Botany*, **95**, 45–90.

Brockington, S.F., Alexandre, R., Ramdial, J., Moore, M.J., Crawley, S., Dhingra, A., Hilu, K., Soltis, D.E. and Soltis, P.S. (2009) Phylogeny of the Caryophyllales Sensu Lato: Revisiting Hypotheses on Pollination Biology and Perianth Differentiation in the Core Caryophyllales. *International Journal of Plant Sciences*, **170**, 627–643.

Brockington, S.F., Alvarez-Fernandez, R., Landis, J.B., Alcorn, K., Walker, R.H., Thomas, M.M., Hileman, L.C. and Glover, B.J. (2013) Evolutionary Analysis of the MIXTA Gene Family Highlights Potential Targets for the Study of Cellular Differentiation. *Molecular Biology and Evolution*, **30**, 526–540.

Brockington, S.F., Rudall, P.J., Frohlich, M.W., Oppenheimer, D.G., Soltis, P.S. and Soltis, D.E. (2012) ‘Living stones’ reveal alternative petal identity programs within the core eudicots. *The Plant Journal*, **69**, 193–203.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.

Coen, E.S. and Meyerowitz, E.M. (1991) The war of the whorls: genetic interactions controlling flower development. *Nature*, **353**, 31–37.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K. and Scheible, W.-R. (2005) Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. *Plant Physiol.*, **139**, 5–17.

Demidenko, N.V., Logacheva, M.D. and Penin, A.A. (2011) Selection and Validation of Reference Genes for Quantitative Real-Time PCR in Buckwheat (Fagopyrum esculentum) Based on Transcriptome Sequence Data L. Mariño-Ramírez, ed. *PLoS ONE*, **6**, e19434.

Dohm, J.C., Minoche, A.E., Holtgräwe, D., et al. (2013) The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). *Nature*, **505**, 546–549.

Doyle, Jeff and Doyle, Jill (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, 11–15.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

Emms, D.M. and Kelly, S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**. Available at: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y [Accessed August 9, 2020].

Fang, X., Zhang, Yuanli, Zhang, Yuke, et al. (2019) *De novo* transcriptome assembly and identification of genes related to seed size in common buckwheat (*Fagopyrum esculentum* M.). *Breed. Sci.*, **69**, 487–497.

Fang, Z., Hou, Z., Wang, S., Liu, Z., Wei, S., Zhang, Y., Song, J. and Yin, J. (2019) Transcriptome Analysis Reveals the Accumulation Mechanism of Anthocyanins in Buckwheat (Fagopyrum esculentum Moench) Cotyledons and Flowers. *IJMS*, **20**, 1493.

Fang, Z., Qi, R., Li, X. and Liu, Z. (2014) Ectopic expression of FaesAP3, a Fagopyrum esculentum (Polygonaceae) AP3 orthologous gene rescues stamen development in an Arabidopsis ap3 mutant. *Gene*, **550**, 200–206.

Fang, Z.-W., Li, X.-P., Li, X.-F. and Liu, Z.-X. (2015) FaesPI, a Fagopyrum esculentum PISTILLATA ortholog, is involved only in stamen development. *J. Plant Biol.*, **58**, 102–109.

Fesenko, A.N. (2009) *New methods of buckwheat breeding [Novye metody selekcii grechihi]*. D. Sci. N.I. Vavilov All-Russian Institute of plant genetic resources.

Fesenko, A.N., Amelin, A.V., Fesenko, I.N., Biryukova, O.V. and Zaikin, V.V. (2018) Novyj sort grechihi Dasha [Dasha - a novel cultivar of buckwheat]. Available at: https://cyberleninka.ru/article/n/novyy-sort-grechihi-dasha/pdf [Accessed May 26, 2020].

Gabaldón, T. and Koonin, E.V. (2013) Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, **14**, 360–366.

Giménez-Bastida, J.A. and Zieliński, H. (2015) Buckwheat as a Functional Food and Its Effects on Health. *Journal of Agricultural and Food Chemistry*, **63**, 7896–7913.

Glover, B.J., Perez-Rodriguez, M. and Martin, C. (1998) Development of several epidermal cell types can be specified by the same MYB-related plant transcription factor. *Development*, **125**, 3497–3508.

Gu, Z., Eils, R. and Schlesner, M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

Guo, X., Hu, Z., Yin, W., Yu, X., Zhu, Z., Zhang, J. and Chen, G. (2016) The tomato floral homeotic protein FBP1-like gene, SlGLO1, plays key roles in petal and stamen development. *Sci Rep*, **6**, 20454.

Hanzawa, Y., Money, T. and Bradley, D. (2005) A single amino acid converts a repressor to an activator of flowering. *Proceedings of the National Academy of Sciences*, **102**, 7748–7753.

Honda, M., Tamura, H., Kimura, T., Kinoshita, T., Mastsufuru, H. and Sasto, T. (2007) Control of Lead Polluted Leachate in a Box-scale Phytoremediation Test using Common Buckwheat (*Fagopyrum esculentum* Moench) grown on Lead Contaminated Soil. *Environmental Technology*, **28**, 425–431.

Hong, S.-P., Craene, L.R. de and Smets, E. (1998) Systematic significance of tepal surface morphology in tribes Persicarieae and Polygoneae (Polygonaceae). *Botanical Journal of the Linnean Society*, **127**, 91–116.

Irish, V.F. and Sussex, I.M. (1990) Function of the apetala-1 gene during Arabidopsis floral development. *Plant Cell*, **2**, 741–753.

Jarvis, D.E., Ho, Y.S., Lightfoot, D.J., et al. (2017) The genome of Chenopodium quinoa. *Nature*, **542**, 307–312.

Joshi, D.C., Chaudhari, G.V., Sood, S., et al. (2019) Revisiting the versatile buckwheat: reinvigorating genetic gains through integrated breeding and genomics approach. *Planta*, **250**, 783–801.

Kajitani, R., Toshimoto, K., Noguchi, H., et al. (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research*, **24**, 1384–1395.

Kanno, A., Saeki, H., Kameya, T., Saedler, H. and Theissen, G. (2003) Heterotopic Expression of Class B Floral Homeotic Genes Supports a Modified ABC Model for Tulip (Tulipa Gesneriana). *Plant Molecular Biology*, **52**, 831–841.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, **16**, 111–120.

Klepikova, A.V., Kasianov, A.S., Gerasimov, E.S., Logacheva, M.D. and Penin, A.A. (2016) A High Resolution Map of the *Arabidopsis thaliana* Developmental Transcriptome Based on RNA-seq Profiling. *The Plant Journal*. Available at: http://doi.wiley.com/10.1111/tpj.13312 [Accessed September 8, 2016].

Lei, G.J., Yokosho, K., Yamaji, N., Fujii-Kashino, M. and Ma, J.F. (2017) Functional characterization of two half-size ABC transporter genes in aluminium-accumulating buckwheat. *New Phytologist*, **215**, 1080–1089.

Leushkin, E.V., Sutormin, R.A., Nabieva, E.R., Penin, A.A., Kondrashov, A.S. and Logacheva, M.D. (2013) The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. *BMC Genomics*, **14**, 476.

Liu, Z., Fei, Y., Zhang, K. and Fang, Z. (2019) Ectopic Expression of a Fagopyrum esculentum APETALA1 Ortholog only Rescues Sepal Development in Arabidopsis ap1 Mutant. *IJMS*, **20**, 2021.

Llorens, C., Futami, R., Covelli, L., et al. (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Research*, **39**, D70–D74.

Logacheva, M.D., Fesenko, I.N., Fesenko, A.N. and Penin, A.A. (2008) Genetic and morphological analysis of floral homeotic mutants *tepal-like bract* and *fagopyrum apetala* of *Fagopyrum esculentum*. *Botany*, **86**, 367–375.

Logacheva, M.D., Kasianov, A.S., Vinogradov, D.V., Samigullin, T.H., Gelfand, M.S., Makeev, V.J. and Penin, A.A. (2011) De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum). *BMC Genomics*, **12**, 30.

Logacheva, M.D., Schelkunov, M.I., Fesenko, A.N., Kasianov, A.S. and Penin, A.A. (2020) Mitochondrial Genome of Fagopyrum esculentum and the Genetic Diversity of Extranuclear Genomes in Buckwheat. *Plants*, **9**, 618.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.

Ma, J. and Bennetzen, J.L. (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences*, **101**, 12404–12410.

Ma, J.F., Zheng, S.J., Matsumoto, H. and Hiradate, S. (1997) Detoxifying aluminium with buckwheat. *Nature*, **390**, 569–570.

Machado, F.B., Moharana, K.C., Almeida-Silva, F., Gazara, R.K., Pedrosa-Silva, F., Coelho, F.S., Grativol, C. and Venancio, T.M. (2020) Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (*Glycine max*) expression atlas. *Plant J*, tpj.14850.

Manchado-Rojo, M., Delgado-Benarroch, L., Roca, M.J., Weiss, J. and Egea-Cortines, M. (2012) Quantitative levels of Deficiens and Globosa during late petal development show a complex transcriptional network topology of B function. *Plant J.*, **72**, 294–307.

Martino, G. de, Pan, I., Emmanuel, E., Levy, A. and Irish, V.F. (2006) Functional Analyses of Two Tomato *APETALA3* Genes Demonstrate Diversification in Their Roles in Regulating Floral Development. *Plant Cell*, **18**, 1833–1845.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Haeseler, A. von and Lanfear, R. (2020) IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era E. Teeling, ed. *Molecular Biology and Evolution*, **37**, 1530–1534.

Murigneux, V., Rai, S.K., Furtado, A., et al. (2020) *Comparison of long read methods for sequencing and assembly of a plant genome*, Genomics. Available at: http://biorxiv.org/lookup/doi/10.1101/2020.03.16.992933 [Accessed May 25, 2020].

Nagano, M., Aii, J., Campbell, C., Kawasaki, S. and Adachi, T. (2000) Genome size analysis of the genus Fagopyrum. *Fagopyrum*, **17**, 35–39.

Nakamura, T., Fukuda, T., Nakano, M., Hasebe, M., Kameya, T. and Kanno, A. (2005) The modified ABC model explains the development of the petaloid perianth of Agapanthus praecox ssp. orientalis (Agapanthaceae) flowers. *Plant Mol Biol*, **58**, 435–445.

Neethirajan, S., Hirose, T., Wakayama, J., Tsukamoto, K., Kanahara, H. and Sugiyama, S. (2011) Karyotype Analysis of Buckwheat Using Atomic Force Microscopy. *Microsc Microanal*, **17**, 572–577.

Novák, P., Ávila Robledillo, L., Koblížková, A., Vrbová, I., Neumann, P. and Macas, J. (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*, **45**, e111–e111.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J. and Macas, J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, **29**, 792–793.

Ober, D. (2005) Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends in Plant Science*, **10**, 444–449.

Ohsako, T. and Li, C. (2020) Classification and systematics of the *Fagopyrum* species. *Breed. Sci.*, **70**, 93–100.

Oshima, Y., Shikata, M., Koyama, T., Ohtsubo, N., Mitsuda, N. and Ohme-Takagi, M. (2013) MIXTA-Like Transcription Factors and WAX INDUCER1/SHINE1 Coordinately Regulate Cuticle Development in *Arabidopsis* and *Torenia fournieri*. *Plant Cell*, **25**, 1609–1624.

Pan, Z.-J., Chen, You-Yi, Du, J.-S., Chen, Yun-Yu, Chung, M.-C., Tsai, W.-C., Wang, C.-N. and Chen, H.-H. (2014) Flower development of Phalaenopsis orchid involves functionally divergent SEPALLATA-like genes. *New Phytol.*, **202**, 1024–1042.

Penin, A.A., Klepikova, A.V., Kasianov, A.S., Gerasimov, E.S. and Logacheva, M.D. (2019) Comparative Analysis of Developmental Transcriptome Maps of Arabidopsis thaliana and Solanum lycopersicum. *Genes*, **10**, 50.

Perez-Rodriguez, M. (2005) Development of three different cell types is associated with the activity of a specific MYB transcription factor in the ventral petal of Antirrhinum majus flowers. *Development*, **132**, 359–370.

Schmidt, M.H.-W., Vogel, A., Denton, A.K., et al. (2017) De Novo Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing. *Plant Cell*, **29**, 2336–2348.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Steinbiss, S., Willhoeft, U., Gremme, G. and Kurtz, S. (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research*, **37**, 7002–7013.

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. and Gao, G. (2019) PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Research*, gkz1020.

Tulpan, D. and Leger, S. (2017) The Plant Orthology Browser: An Orthology and Gene-Order Visualizer for Plant Comparative Genomics. *The Plant Genome*, **10**, plantgenome2016.08.0078.

van Tunen, Arjen, Eikeboom, W and Angenent, Gerko (1993) Floral organogenesis in Tulipa. *Flowering Newsletters*, 33–37.

Vitte, C. and Panaud, O. (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet Genome Res*, **110**, 91–107.

Whitney, H.M., Bennett, K.M.V., Dorling, M., Sandbach, L., Prince, D., Chittka, L. and Glover, B.J. (2011) Why do so many petals have conical epidermal cells? *Annals of Botany*, **108**, 609–616.

Wickham, H. (2009) *Ggplot2: elegant graphics for data analysis*, New York: Springer.

Xu, G.-C., Xu, T.-J., Zhu, R., Zhang, Y., Li, S.-Q., Wang, H.-W. and Li, J.-T. (2019) LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience*, **8**.

Available at: https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giy157/5256637 [Accessed August 5, 2020].

Xu, J.M., Fan, W., Jin, J.F., Lou, H.Q., Chen, W.W., Yang, J.L. and Zheng, S.J. (2017) Transcriptome Analysis of Al-Induced Genes in Buckwheat (Fagopyrum esculentum Moench) Root Apex: New Insight into Al Toxicity and Resistance Mechanisms in an Al Accumulating Species. *Front. Plant Sci.*, **8**, 1141.

Yamaguchi, T., Lee, D.Y., Miyao, A., Hirochika, H., An, G. and Hirano, H.-Y. (2006) Functional Diversification of the Two C-Class MADS Box Genes *OSMADS3* and *OSMADS58* in *Oryza sativa*. *Plant Cell*, **18**, 15–28.

Yao, G., Jin, J.-J., Li, H.-T., et al. (2019) Plastid phylogenomic insights into the evolution of Caryophyllales. *Molecular Phylogenetics and Evolution*, **134**, 74–86.

Yasui, Y., Hirakawa, H., Ueno, M., Matsui, K., Katsube-Tanaka, T., Yang, S.J., Aii, J., Sato, S. and Mori, M. (2016) Assembly of the draft genome of buckwheat and its applications in identifying agronomically useful genes. *DNA Research*, **23**, 215–224.

Zhang, L., Li, X., Ma, B., et al. (2017) The Tartary Buckwheat Genome Provides Insights into Rutin Biosynthesis and Abiotic Stress Tolerance. *Molecular Plant*, **10**, 1224–1237.

Zhang, R.-G., Wang, Z.-X., Ou, S. and Li, G.-Y. (2019) *TEsorter: lineage-level classification of transposable elements using conserved protein domains*, Bioinformatics. Available at: http://biorxiv.org/lookup/doi/10.1101/800177 [Accessed September 9, 2020].

Zhou, Z., Cong, P., Tian, Y. and Zhu, Y. (2017) Using RNA-seq data to select reference genes for normalizing gene expression in apple roots Y. Huang, ed. *PLoS ONE*, **12**, e0185288.

Figure 1.TIF

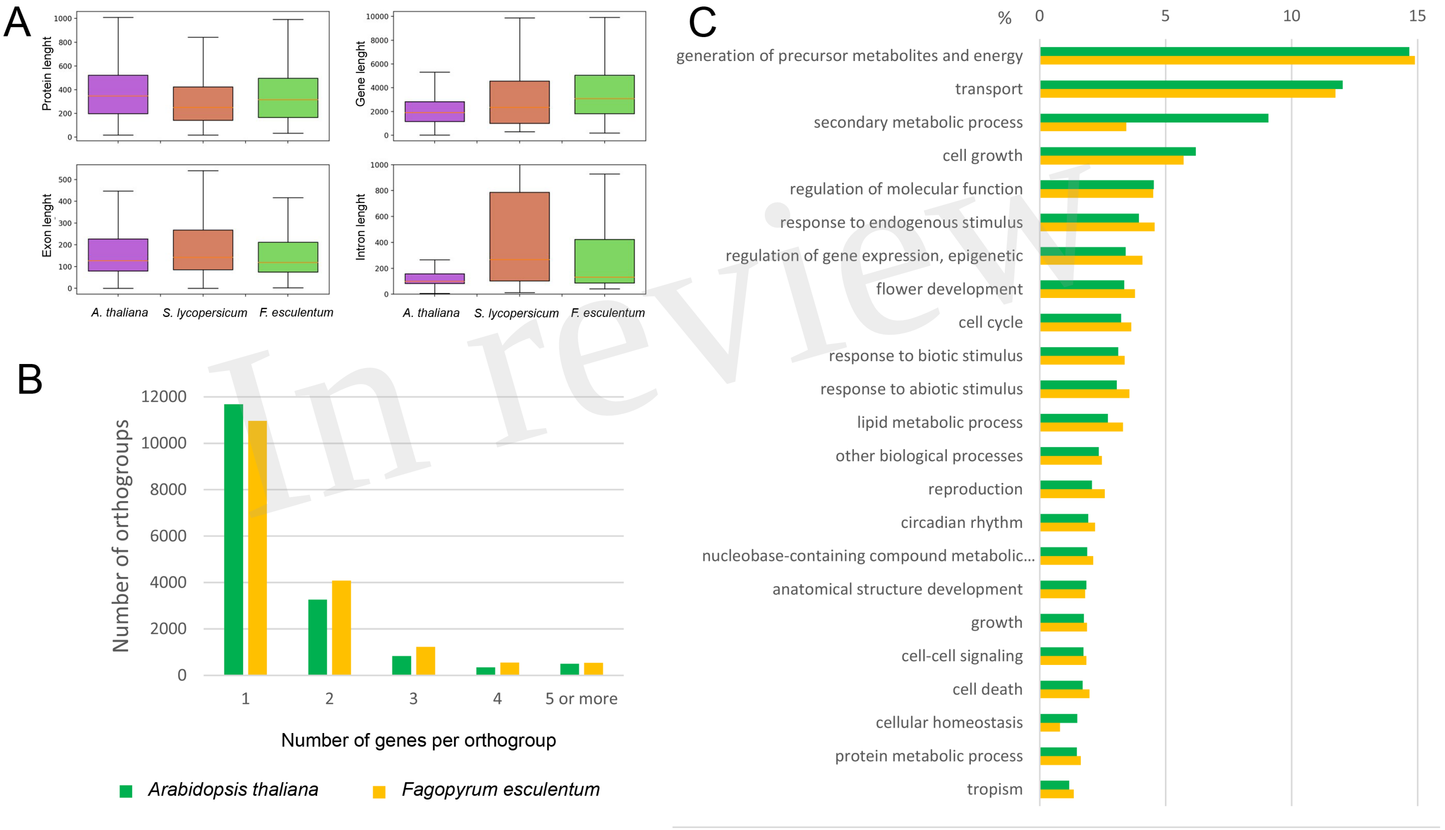

A
Protein lenght
1000
800
600
400
200
0
Gene lenght
10000
8000
6000
4000
2000
0
Exon lenght
500
400
300
200
100
0
Intron lenght
1000
800
600
400
200
0
A. thaliana
S. lycopersicum
F. esculentum
A. thaliana
S. lycopersicum
F. esculentum
B
Number of orthogroups
12000
10000
8000
6000
4000
2000
0
1
2
3
4
5 or more
Number of genes per orthogroup
🟩 Arabidopsis thaliana
🟨 Fagopyrum esculentum
C
%
0
5
10
15
generation of precursor metabolites and energy
transport
secondary metabolic process
cell growth
regulation of molecular function
response to endogenous stimulus
regulation of gene expression, epigenetic
flower development
cell cycle
response to biotic stimulus
response to abiotic stimulus
lipid metabolic process
other biological processes
reproduction
circadian rhythm
nucleobase-containing compound metabolic...
anatomical structure development
growth
cell-cell signaling
cell death
cellular homeostasis
protein metabolic process
tropism

Figure 2.JPEG

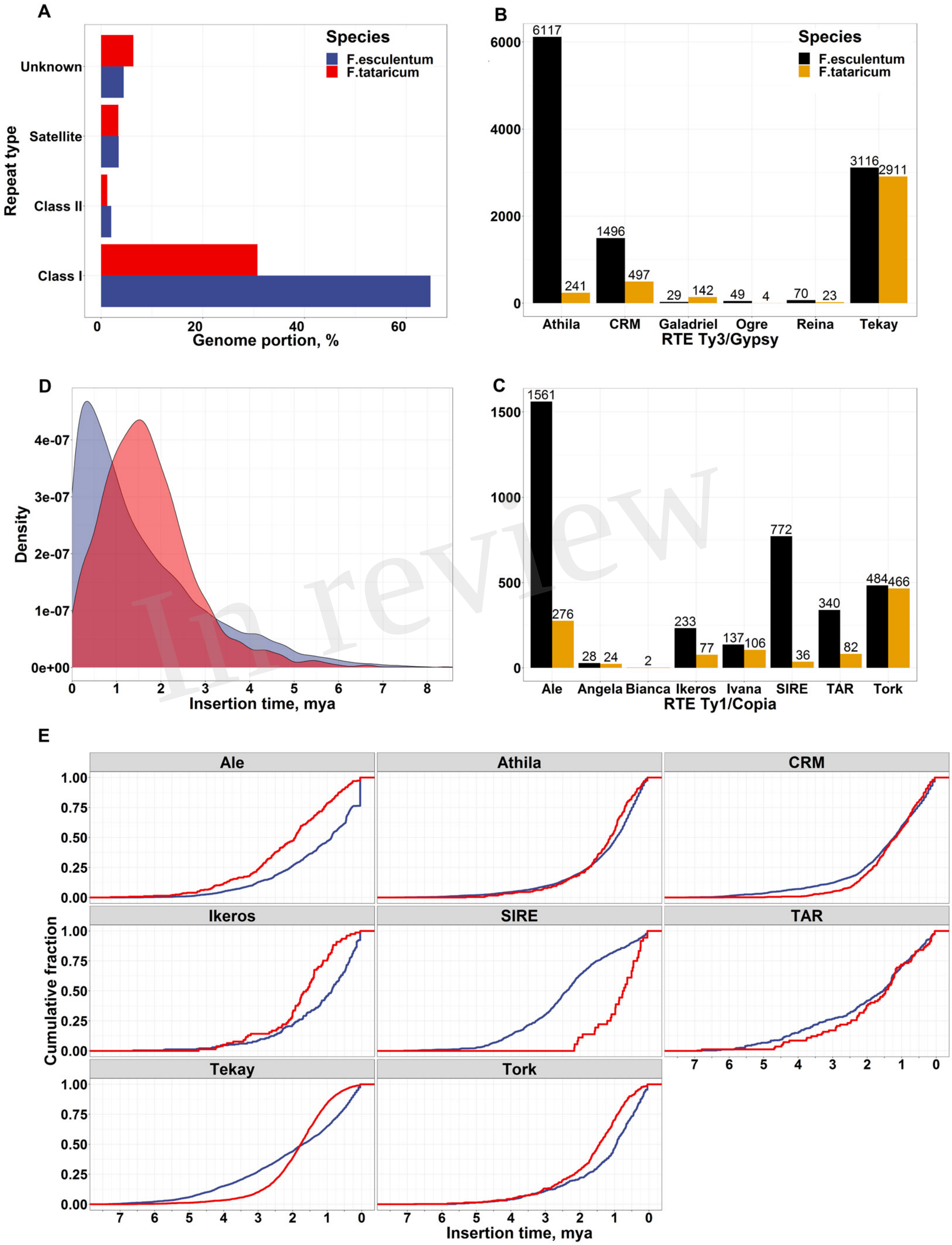
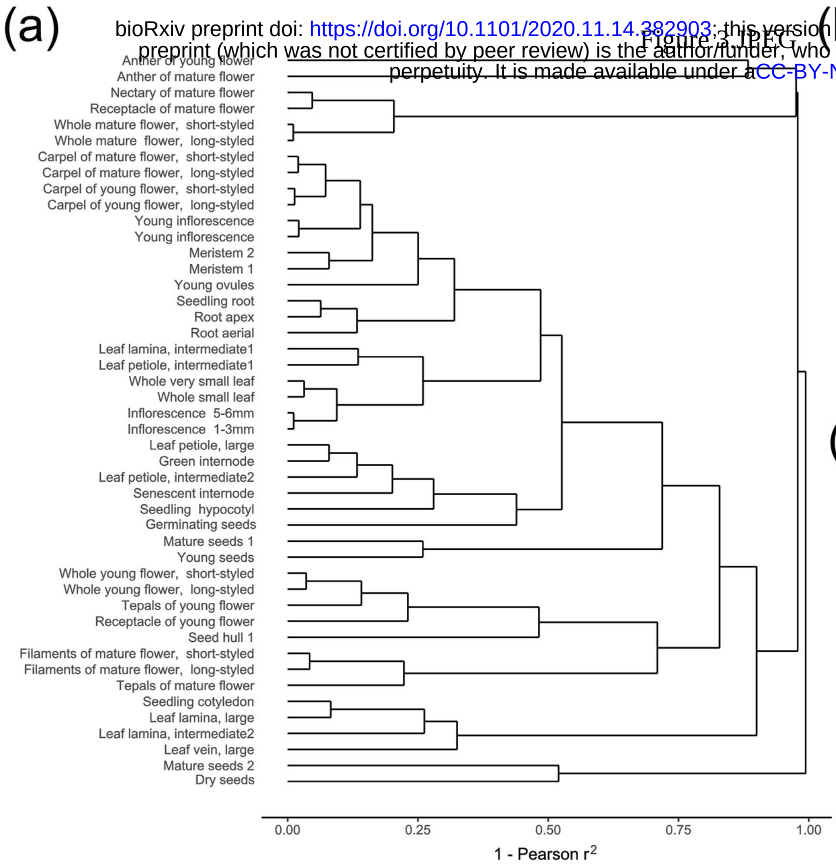
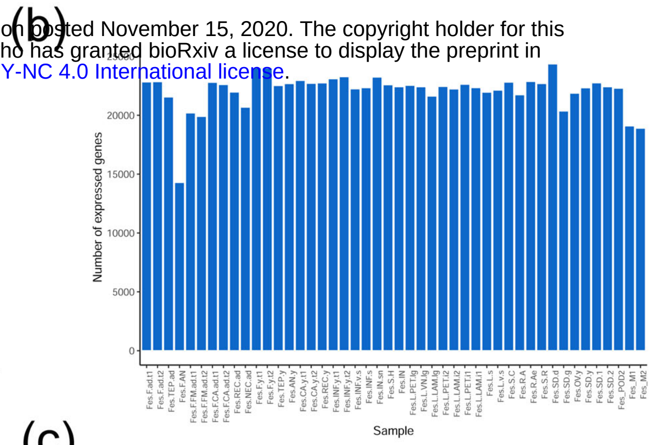
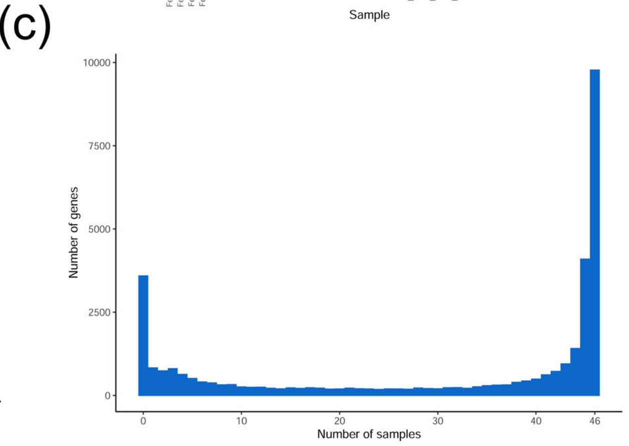
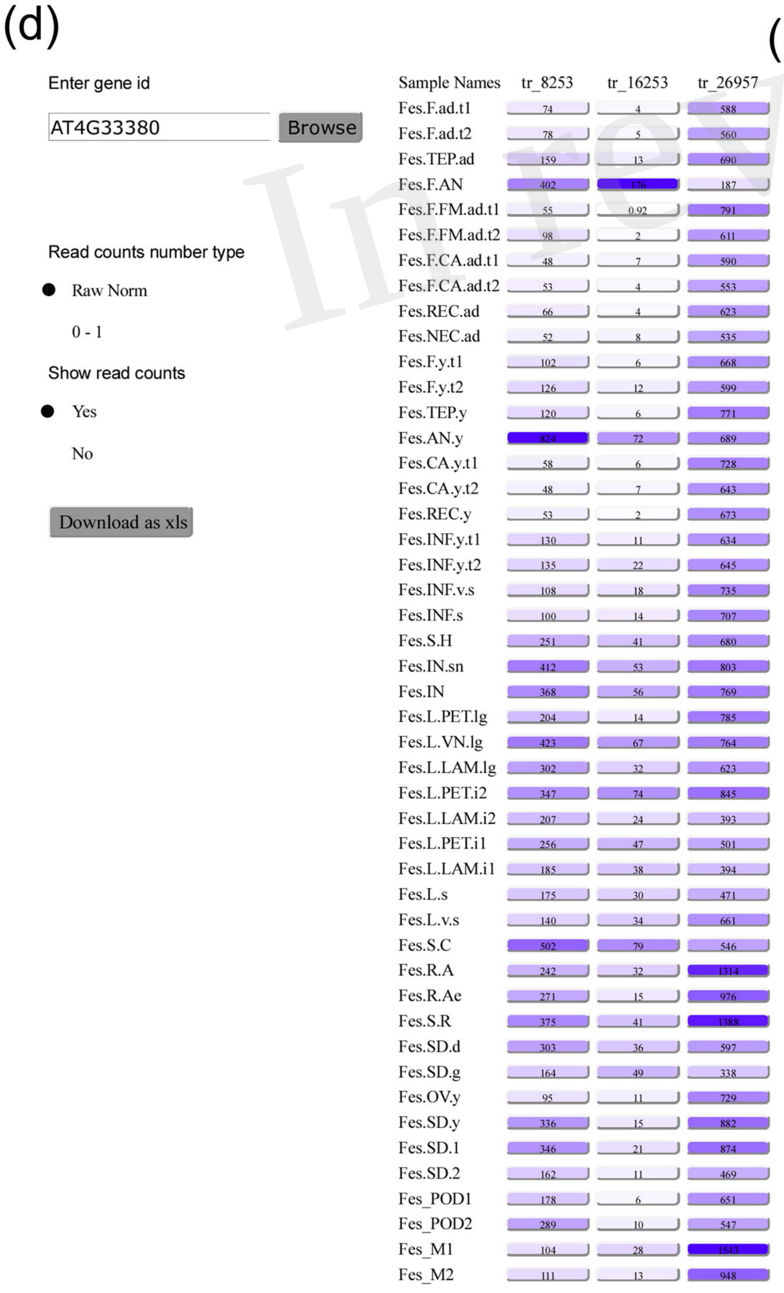


Figure 3.JPEG









Enter gene id

AT4G33380 Browse

Read counts number type

● Raw Norm

0 - 1

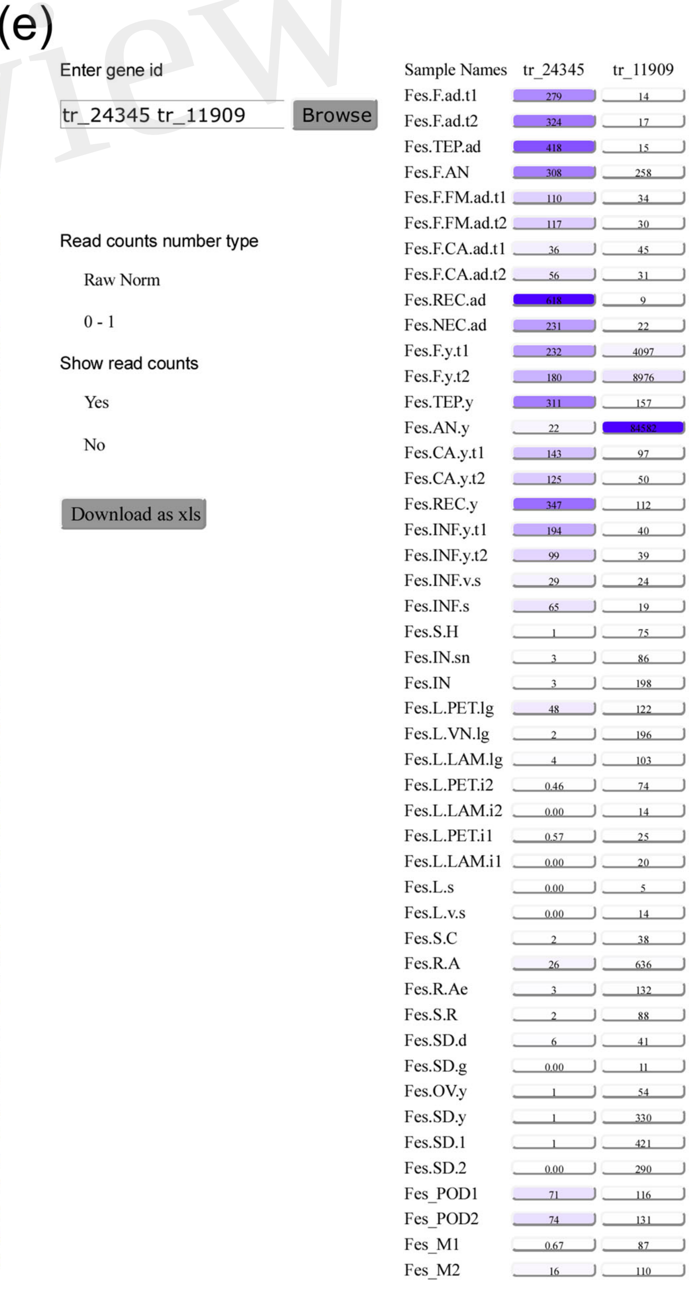
Show read counts

● Yes

No

Download as xls

Sample Names	tr_8253	tr_16253	tr_26957
Fes.F.ad.t1	74	4	588
Fes.F.ad.t2	78	5	560
Fes.TEP.ad	159	13	690
Fes.F.AN	402	176	187
Fes.F.FM.ad.t1	55	0.92	791
Fes.F.FM.ad.t2	98	2	611
Fes.F.CA.ad.t1	48	7	590
Fes.F.CA.ad.t2	53	4	553
Fes.REC.ad	66	4	623
Fes.NEC.ad	52	8	535
Fes.F.y.t1	102	6	668
Fes.F.y.t2	126	12	599
Fes.TEP.y	120	6	771
Fes.AN.y	824	72	689
Fes.CA.y.t1	58	6	728
Fes.CA.y.t2	48	7	643
Fes.REC.y	53	2	673
Fes.INF.y.t1	130	11	634
Fes.INF.y.t2	135	22	645
Fes.INF.v.s	108	18	735
Fes.INF.s	100	14	707
Fes.S.H	251	41	680
Fes.IN.sn	412	53	803
Fes.IN	368	56	769
Fes.L.PET.lg	204	14	785
Fes.L.VN.lg	423	67	764
Fes.L.LAM.lg	302	32	623
Fes.L.PET.i2	347	74	845
Fes.L.LAM.i2	207	24	393
Fes.L.PET.i1	256	47	501
Fes.L.LAM.i1	185	38	394
Fes.L.s	175	30	471
Fes.L.v.s	140	34	661
Fes.S.C	502	79	546
Fes.R.A	242	32	1314
Fes.R.Ae	271	15	976
Fes.S.R	375	41	1388
Fes.SD.d	303	36	597
Fes.SD.g	164	49	338
Fes.OV.y	95	11	729
Fes.SD.y	336	15	882
Fes.SD.1	346	21	874
Fes.SD.2	162	11	469
Fes_POD1	178	6	651
Fes_POD2	289	10	547
Fes_M1	104	28	1543
Fes_M2	111	13	948



Enter gene id

tr_24345 tr_11909 Browse

Read counts number type

Raw Norm

0 - 1

Show read counts

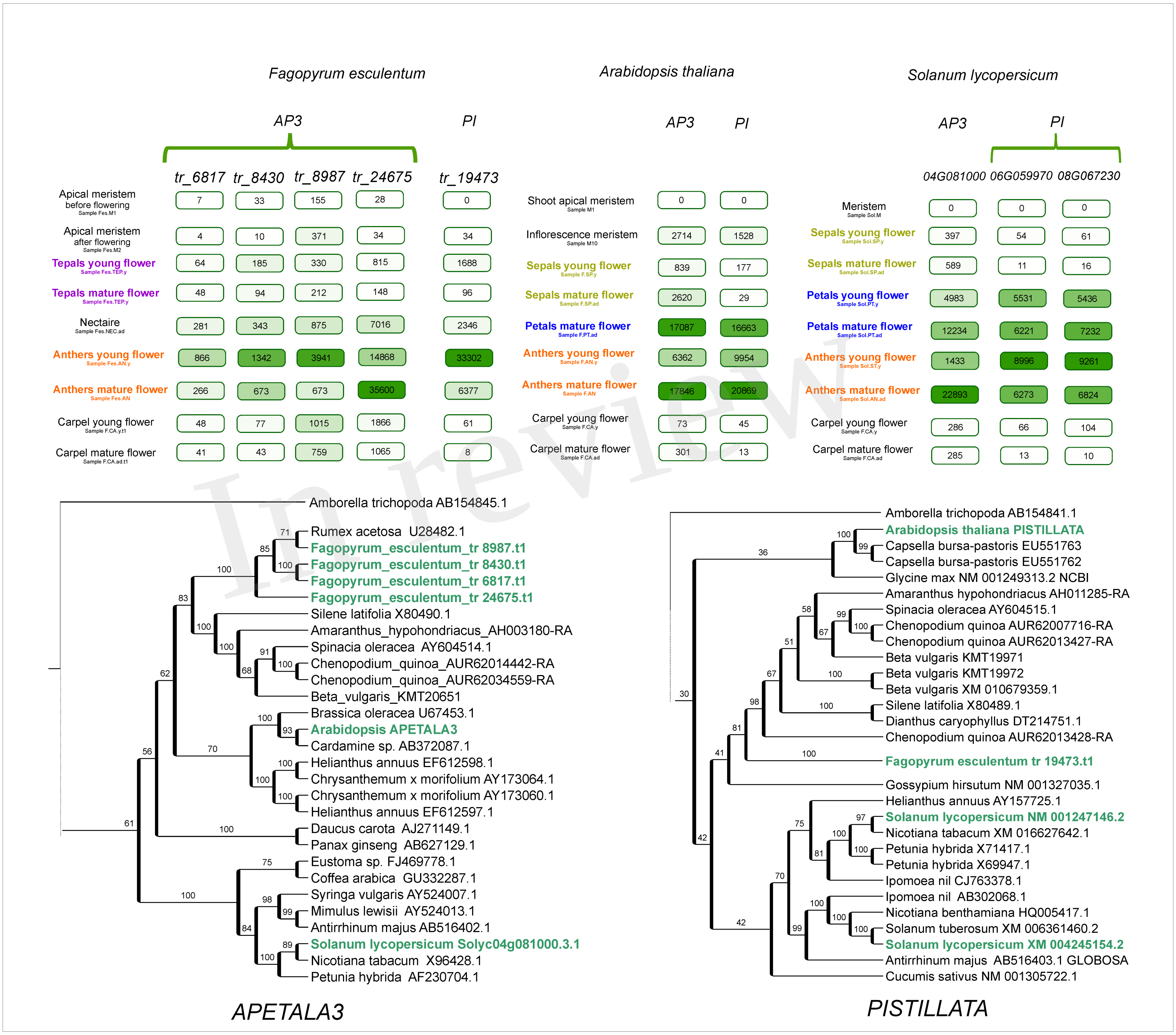
Yes

No

Download as xls

Sample Names	tr_24345	tr_11909
Fes.F.ad.t1	279	14
Fes.F.ad.t2	324	17
Fes.TEP.ad	418	15
Fes.F.AN	308	258
Fes.F.FM.ad.t1	110	34
Fes.F.FM.ad.t2	117	30
Fes.F.CA.ad.t1	36	45
Fes.F.CA.ad.t2	56	31
Fes.REC.ad	618	9
Fes.NEC.ad	231	22
Fes.F.y.t1	232	4097
Fes.F.y.t2	180	8976
Fes.TEP.y	311	157
Fes.AN.y	22	84582
Fes.CA.y.t1	143	97
Fes.CA.y.t2	125	50
Fes.REC.y	347	112
Fes.INF.y.t1	194	40
Fes.INF.y.t2	99	39
Fes.INF.v.s	29	24
Fes.INF.s	65	19
Fes.S.H	1	75
Fes.IN.sn	3	86
Fes.IN	3	198
Fes.L.PET.lg	48	122
Fes.L.VN.lg	2	196
Fes.L.LAM.lg	4	103
Fes.L.PET.i2	0.46	74
Fes.L.LAM.i2	0.00	14
Fes.L.PET.i1	0.57	25
Fes.L.LAM.i1	0.00	20
Fes.L.s	0.00	5
Fes.L.v.s	0.00	14
Fes.S.C	2	38
Fes.R.A	26	636
Fes.R.Ae	3	132
Fes.S.R	2	88
Fes.SD.d	6	41
Fes.SD.g	0.00	11
Fes.OV.y	1	54
Fes.SD.y	1	330
Fes.SD.1	1	421
Fes.SD.2	0.00	290
Fes_POD1	71	116
Fes_POD2	74	131
Fes_M1	0.67	87
Fes_M2	16	110

Figure 4.TIF



Fagopyrum esculentum

	AP3				*PI*
	tr_6817	*tr_8430*	*tr_8987*	*tr_24675*	*tr_19473*
Apical meristem before flowering Sample Fes.M1	7	33	155	28	0
Apical meristem after flowering Sample Fes.M2	4	10	371	34	34
Tepals young flower Sample Fes.TEP.y	64	185	330	815	1688
Tepals mature flower Sample Fes.TEP.y	48	94	212	148	96
Nectaire Sample Fes.NEC.ad	281	343	875	7016	2346
Anthers young flower Sample Fes.AN.y	866	1342	3941	14868	33302
Anthers mature flower Sample Fes.AN	266	673	673	35600	6377
Carpel young flower Sample F.CA.y.t1	48	77	1015	1866	61
Carpel mature flower Sample F.CA.ad.t1	41	43	759	1065	8

Arabidopsis thaliana

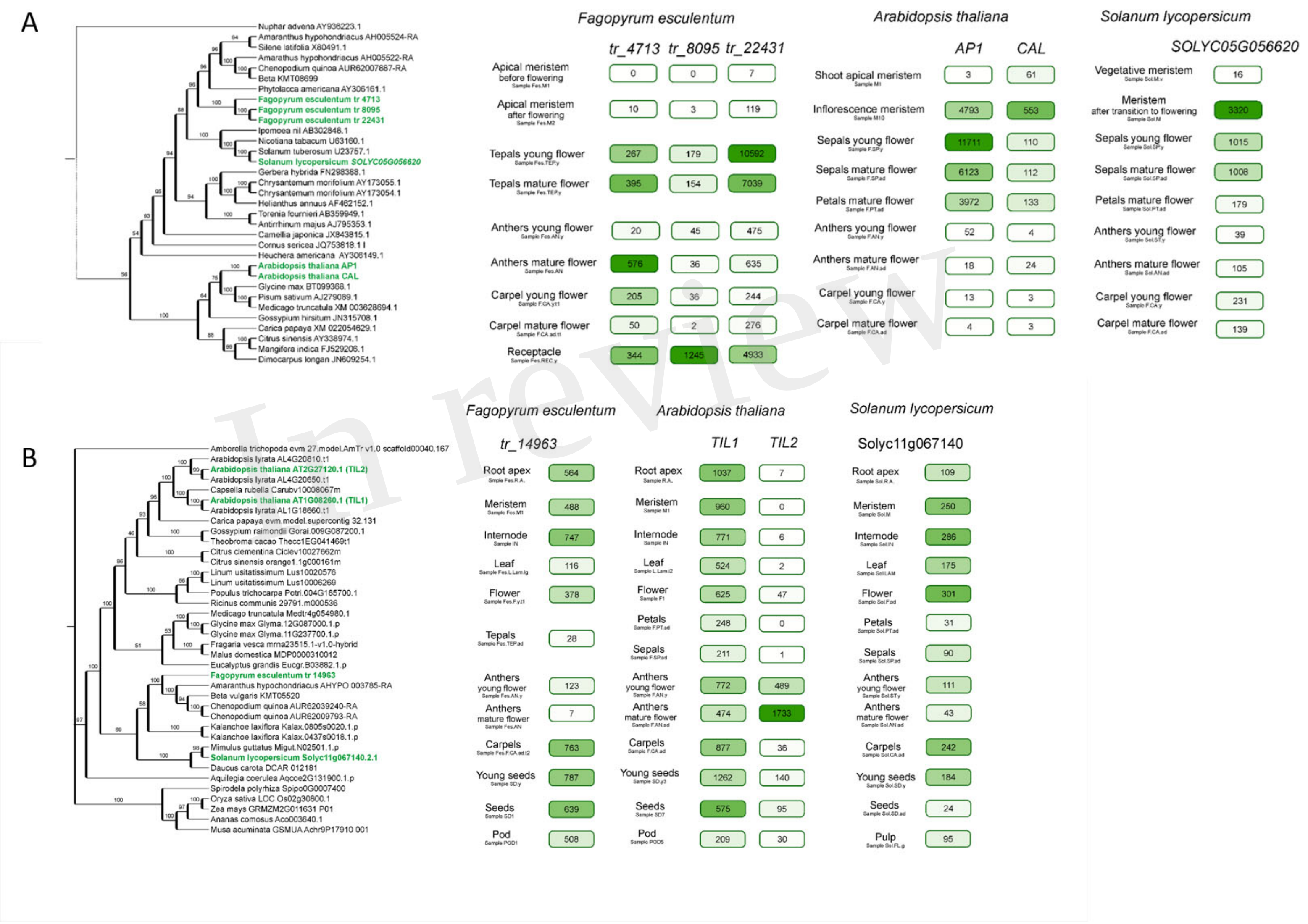
	AP3	*PI*
Shoot apical meristem Sample M1	0	0
Inflorescence meristem Sample M10	2714	1528
Sepals young flower Sample F.SP.y	839	177
Sepals mature flower Sample F.SP.ad	2620	29
Petals mature flower Sample F.PT.ad	17087	16663
Anthers young flower Sample F.AN.y	6362	9954
Anthers mature flower Sample F.AN	17846	20869
Carpel young flower Sample F.CA.y	73	45
Carpel mature flower Sample F.CA.ad	301	13

Solanum lycopersicum

	AP3	*PI*	
	04G081000	*06G059970*	*08G067230*
Meristem Sample Sol.M	0	0	0
Sepals young flower Sample Sol.SP.y	397	54	61
Sepals mature flower Sample Sol.SP.ad	589	11	16
Petals young flower Sample Sol.PT.y	4983	5531	5436
Petals mature flower Sample Sol.PT.ad	12234	6221	7232
Anthers young flower Sample Sol.ST.y	1433	8996	9261
Anthers mature flower Sample Sol.AN.ad	22893	6273	6824
Carpel young flower Sample F.CA.y	286	66	104
Carpel mature flower Sample F.CA.ad	285	13	10

Amborella trichopoda AB154845.1
Rumex acetosa U28482.1
Fagopyrum_esculentum_tr 8987.t1
Fagopyrum_esculentum_tr 8430.t1
Fagopyrum_esculentum_tr 6817.t1
Fagopyrum_esculentum_tr 24675.t1
Silene latifolia X80490.1
Amaranthus_hypohondriacus_AH003180-RA
Spinacia oleracea AY604514.1
Chenopodium_quinoa_AUR62014442-RA
Chenopodium_quinoa_AUR62034559-RA
Beta_vulgaris_KMT20651
Brassica oleracea U67453.1
Arabidopsis APETALA3
Cardamine sp. AB372087.1
Helianthus annuus EF612598.1
Chrysanthemum x morifolium AY173064.1
Chrysanthemum x morifolium AY173060.1
Helianthus annuus EF612597.1
Daucus carota AJ271149.1
Panax ginseng AB627129.1
Eustoma sp. FJ469778.1
Coffea arabica GU332287.1
Syringa vulgaris AY524007.1
Mimulus lewisii AY524013.1
Antirrhinum majus AB516402.1
Solanum lycopersicum Solyc04g081000.3.1
Nicotiana tabacum X96428.1
Petunia hybrida AF230704.1
71
85
100
100
83
100
100
91
68
100
62
100
93
70
100
100
100
56
61
100
75
100
98
99
84
89
100
APETALA3
Amborella trichopoda AB154841.1
Arabidopsis thaliana PISTILLATA
Capsella bursa-pastoris EU551763
Capsella bursa-pastoris EU551762
Glycine max NM 001249313.2 NCBI
Amaranthus hypohondriacus AH011285-RA
Spinacia oleracea AY604515.1
Chenopodium quinoa AUR62007716-RA
Chenopodium quinoa AUR62013427-RA
Beta vulgaris KMT19971
Beta vulgaris KMT19972
Beta vulgaris XM 010679359.1
Silene latifolia X80489.1
Dianthus caryophyllus DT214751.1
Chenopodium quinoa AUR62013428-RA
Fagopyrum esculentum tr 19473.t1
Gossypium hirsutum NM 001327035.1
Helianthus annuus AY157725.1
Solanum lycopersicum NM 001247146.2
Nicotiana tabacum XM 016627642.1
Petunia hybrida X71417.1
Petunia hybrida X69947.1
Ipomoea nil CJ763378.1
Ipomoea nil AB302068.1
Nicotiana benthamiana HQ005417.1
Solanum tuberosum XM 006361460.2
Solanum lycopersicum XM 004245154.2
Antirrhinum majus AB516403.1 GLOBOSA
Cucumis sativus NM 001305722.1
100
99
36
58
99
67
100
51
67
100
98
100
30
81
41
100
97
75
100
81
100
42
70
100
100
42
99
100
PISTILLATA

Figure 5.JPEG


A
Fagopyrum esculentum
tr_4713 tr_8095 tr_22431
Apical meristem before flowering 0 0 7
Apical meristem after flowering 10 3 119
Tepals young flower 267 179 10592
Tepals mature flower 395 154 7039
Anthers young flower 20 45 475
Anthers mature flower 576 36 635
Carpel young flower 205 36 244
Carpel mature flower 50 2 276
Receptacle 344 1245 4933
Arabidopsis thaliana
AP1 CAL
Shoot apical meristem 3 61
Inflorescence meristem 4793 553
Sepals young flower 11711 110
Sepals mature flower 6123 112
Petals mature flower 3972 133
Anthers young flower 52 4
Anthers mature flower 18 24
Carpel young flower 13 3
Carpel mature flower 4 3
Solanum lycopersicum
SOLYC05G056620
Vegetative meristem 16
Meristem after transition to flowering 3320
Sepals young flower 1015
Sepals mature flower 1008
Petals mature flower 179
Anthers young flower 39
Anthers mature flower 105
Carpel young flower 231
Carpel mature flower 139
Fagopyrum esculentum tr 4713
Fagopyrum esculentum tr 8095
Fagopyrum esculentum tr 22431
Solanum lycopersicum SOLYC05G056620
Arabidopsis thaliana AP1
Arabidopsis thaliana CAL
B
Fagopyrum esculentum
tr_14963
Root apex 564
Meristem 488
Internode 747
Leaf 116
Flower 378
Tepals 28
Anthers young flower 123
Anthers mature flower 7
Carpels 763
Young seeds 787
Seeds 639
Pod 508
Arabidopsis thaliana
TIL1 TIL2
Root apex 1037 7
Meristem 960 0
Internode 771 6
Leaf 524 2
Flower 625 47
Petals 248 0
Sepals 211 1
Anthers young flower 772 489
Anthers mature flower 474 1733
Carpels 877 36
Young seeds 1262 140
Seeds 575 95
Pod 209 30
Solanum lycopersicum
Solyc11g067140
Root apex 109
Meristem 250
Internode 286
Leaf 175
Flower 301
Petals 31
Sepals 90
Anthers young flower 111
Anthers mature flower 43
Carpels 242
Young seeds 184
Seeds 24
Pulp 95
Arabidopsis thaliana AT2G27120.1 (TIL2)
Arabidopsis thaliana AT1G08260.1 (TIL1)
Fagopyrum esculentum tr 14963
Solanum lycopersicum Solyc11g067140.2.1