De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum)

Logacheva et al.
De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum)

Maria D Logacheva1,2,4*, Artem S Kasianov3, Dmitriy V Vinogradov4, Tagir H Samigullin1, Mikhail S Gelfand4,5, Vsevolod J Makeev3,6,7, Aleksey A Penin2,4,8

Abstract

Background: Transcriptome sequencing data has become an integral component of modern genetics, genomics and evolutionary biology. However, despite advances in the technologies of DNA sequencing, such data are lacking for many groups of living organisms, in particular, many plant taxa. We present here the results of transcriptome sequencing for two closely related plant species. These species, Fagopyrum esculentum and F. tataricum, belong to the order Caryophyllales - a large group of flowering plants with uncertain evolutionary relationships. F. esculentum (common buckwheat) is also an important food crop. Despite these practical and evolutionary considerations Fagopyrum species have not been the subject of large-scale sequencing projects.

Results: Normalized cDNA corresponding to genes expressed in flowers and inflorescences of F. esculentum and F. tataricum was sequenced using the 454 pyrosequencing technology. This resulted in 267 (for F. esculentum) and 229 (F. tataricum) thousands of reads with average length of 341-349 nucleotides. De novo assembly of the reads produced about 25 thousands of contigs for each species, with 7.5-8.2× coverage. Comparative analysis of two transcriptomes demonstrated their overall similarity but also revealed genes that are presumably differentially expressed. Among them are retrotransposon genes and genes involved in sugar biosynthesis and metabolism. Thirteen single-copy genes were used for phylogenetic analysis; the resulting trees are largely consistent with those inferred from multigenic plastid datasets. The sister relationships of the Caryophyllales and asterids now gained high support from nuclear gene sequences.

Conclusions: 454 transcriptome sequencing and de novo assembly was performed for two congeneric flowering plant species, F. esculentum and F. tataricum. As a result, a large set of cDNA sequences that represent orthologs of known plant genes as well as potential new genes was generated.

Background

Transcriptome sequencing is a convenient way to rapidly obtain information on the expressed fraction of genome. With the advent of next-generation sequencing transcriptomic data for many species became available. However many important taxonomic groups of living organisms remain underrepresented. This is especially pressing problem for plants since few of them have been a priority for large-scale sequencing projects. The present study aimed at filling the gap in the genomic data for the genus Fagopyrum, a group of plants important for both practical and theoretical reasons.

Fagopyrum (buckwheat) belongs to the eudicot family Polygonaceae. The genus Fagopyrum comprises about 17 species; one of the species, F. esculentum (common buckwheat), is a crop and honey-producing plant widely cultivated in several countries, in particular, Canada, China, Japan, Russia and Ukraine. Recently, an extensive search for the wild ancestor of the cultivated buckwheat identified the ancestral group, called F. esculentum subspecies ancestrale [1,2]. Two more distant relatives of the cultivated buckwheat are F. homotropicum and F. tataricum [3]. Both of them have an important favorable
trait absent in the common buckwheat, the ability to self-pollination (self-compatibility). Thus they have a potential for use in breeding for the development of self-compatible cultivars.

Despite its economic importance, molecular studies on buckwheat are few. They are mainly confined to the development of molecular markers based on anonymous DNA sequences (RAPD, SSR, AFLP) and the characterization of several proteins [4-6]. Also, the molecular systematics of the genus *Fagopyrum* was extensively studied using nuclear and plastid sequences [7-9], and a complete sequence of the chloroplast genome of the common buckwheat was reported recently [10]. These data, however, have only limited applicability for the buckwheat breeding and improvement. The identification of genes expressed during flower development can enable the search for candidate genes responsible for agriculturally important traits.

The utility of large-scale buckwheat gene sequencing data is not limited to their potential practical applications. They will also contribute to the plant comparative and evolutionary genomics. Not only the buckwheat, but the entire family Polygonaceae was out of scope of molecular genetic studies. The Polygonaceae is a middle-sized family (ca. 1000 species) characterized by several morphological features that are not typical for other eudicots. Among them is an unusual floral structure with flowers having a uniseriate perianth not differentiated into sepals and petals. This feature, as well as the trimerous perianth found in several Polygonaceae genera and considered as ancestral state for the family, is common for monocots and basal angiosperms but rarely found in eudicots.

The genetic control of the floral development has been well studied in several model plants, primarily *Arabidopsis thaliana* and some gene interactions have been found to be conserved in a variety of species [11]. However, it is widely accepted that for understanding of the plant morphological evolution new model systems representing different lineages of the angiosperms should be selected [12,13]. Given the morphological peculiarities of Polygonaceae, *Fagopyrum*, and specifically, *F. esculentum*, a fast growing annual weed with high seed production, is a plausible candidate for such a system.

Many factors are important for species to be useful as a model system, a crucial one being the availability of genomic and transcriptomic data that enable efficient analysis of gene structure, expression and evolution. Until recently, such data were available only for major model species such as *A. thaliana* and rice, and for several species of basal angiosperms [14]. We report characterization of genes expressed in the flower and inflorescence of two species of buckwheat, the common buckwheat *F. esculentum* and *F. tataricum*. This choice was based on the reasons outlined above. One is the economical importance of *F. esculentum* and *F. tataricum* being a potential donor of agriculturally important traits such as self-compatibility and resistance to environmental stresses [15] to the common buckwheat. The other stems from the evolutionary developmental genetics (evo-devo) considerations. These two species have contrasting floral morphology: *F. esculentum* has petaloid tepals and *F. tataricum* has sepaloid ones (Figure 1). Further qualitative and quantitative analysis based on the transcriptome data may help to reveal the mechanisms that are responsible for the morphological difference between these species.

We expect that characterization of buckwheat transcriptomes will contribute not only to the genetics of this genus. The order Caryophyllales is characterized by a striking diversity of the floral structure and the perianth evolution that can hardly be explained based only on the knowledge about classic model species [16]. Few transcriptome datasets of the Caryophyllales are available to date [17,18]. Publicly available, annotated set of buckwheat genes expressed in the flower and inflorescence resulting from this study will facilitate comparative and evolutionary studies within the Caryophyllales, important for the plant evo-devo field.

**Results**

**454 sequencing and assembly**

Pyrosequencing of normalized cDNA libraries resulted in 266782 reads for *F. esculentum* and 229031 for *F. tataricum* with the average length of 349 and 341 nucleotides respectively. The raw data were deposited in the NCBI Sequence Read Archive (SRA) under the accession number SRA023408.

After assembly, 25435 and 25401 contigs were obtained for *F. esculentum* and *F. tataricum*, respectively, while 56874 and 42913 reads were retained as
An automatic ORF prediction revealed open reading frames with length more than 90 bp in about 98% contigs in both buckwheat species. This length, though being much less than the average length of protein-coding sequence in plants, is higher than expected to arise by chance (assessed by random shuffling of the contigs with subsequent ORF prediction) suggesting that most of the contigs correspond to protein-coding genes. This is consistent with the method used for cDNA synthesis, as it is based on the presence of polyA-tail on the 3’-end of most mRNAs, the library is highly enriched in mRNA sequences. Interestingly, most contigs that had no BLAST matches, also contained predicted ORFs with length exceeding the one expected for a random sequence of a similar GC-content (data not shown).

To obtain a functional annotation of the buckwheat genes, we used Gene Ontology (GO, [24]). For both species of buckwheat, about 60% contigs were annotated. This fraction is similar to or exceeds the annotated fraction in other recently published 454-derived transcriptomes, for example those of pine [23] and ginseng [25]. The distribution of GO categories is very similar in *F. esculentum* and *F. tataricum*, with no categories showing significant differences between the species (Figure 3).

The contigs that had no BLASTX hits but contained an ORF with a length exceeding 90 nucleotides were searched against the Pfam database (e-value threshold 1). A similar number of translated contigs had significant match in the database for both species, 1795 for *F. esculentum* and 1775 for *F. tataricum*, with 1108 and 1156, respectively, having a match in pfamA families. Among the latter, about 50% (47% for *F. esculentum* and 50% for *F. tataricum*) had hits to families containing at least one protein from plants.

Since one of our aims was to identify genes that are responsible for the flower development, we specifically searched for buckwheat orthologs of genes and gene families that are known to be involved in the flower development in *Arabidopsis*. Candidate orthologs were identified as bidirectional best BLAST hits. Such orthologs were found (in either *F. esculentum* or *F. tataricum*) for 61 out of 141 genes of *Arabidopsis*. Among them were all genes controlling the floral organ identity (ABC model genes) and some other MADS-box genes; genes controlling stem cell maintenance (*CLAVATA1*, *CLAVATA2* and *WUSCHEL*); negative regulators of *AGAMOUS* (*APETALA2*, *LEUNIG*, *SEIUS*, *BELLRINGER*); genes of the *YABBY* family; auxin response factors and some others (Table 3).

Since this type of search is not able to detect duplications, we performed a manual inspection of the results of BLAST search for several gene families. For the B-class MADS-box gene, *APETALA3* (*AP3*), we were able to find two highly similar sequences. The similarity between them is lower than that expected for alleles, with multiple indels and non-synonymous substitutions. Thus we assume that they represent paralogs arisen

---

**Table 1 Characteristics of raw data and contigs**

|                           | Fagopyrum esculentum | Fagopyrum tataricum |
|---------------------------|----------------------|----------------------|
| Number of reads           | 266782               | 229031               |
| Length, mean (min-max)    | 349 (40-971)         | 340 (40-976)         |
| Number of contigs         | 25435                | 25401                |
| Length, mean (min-max)    | 6984 (42-3607)       | 703 (46-3298)        |
| Number of reads per contig, mean (min-max) | 8.2 (2-224) | 7.5 (2-295) |
from a recent duplication. In another species from the same family, *Rumex acetosa*, two genes with high similarity to AP3 were also found [26]. Hence, this duplication likely occurred in the common ancestor of the Polygonaceae.

**Comparative analysis of the buckwheat transcriptomes**

The contig assembly and annotation characteristics, as well as the raw data characteristics, are similar for *F. esculentum* and *F. tataricum*. Moreover, the data indicate that there is a strong correlation between the

---

**Figure 2 De novo sequencing and assembly characteristics**

Distribution of contig lengths (a), distribution of the number of reads per contig (b).

|        | Fagopyrum esculentum | Fagopyrum tataricum |
|--------|----------------------|---------------------|
| a      |                      |                     |
| b      |                      |                     |
number of reads constituting contigs for orthologous
genes of *F. esculentum* and *F. tataricum* (Figure 4). If a
contig of a certain gene is overlapping with a certain
number of reads, there is a high probability that the
contig of its ortholog from the other species is overlap-
ping with a close number of reads.

We addressed the question whether these transcrip-
tomes completely consist of orthologous genes or there
are genes expressed in one species but with orthologs
that are not represented (absent or not expressed) in the
other species. Such transcripts are plausible candidates
for the determinants of morphological differences
between the species.

To identify these putative differentially expressed genes
we first performed a bidirectional BLAST search of the *F.
esculentum* and *F. tataricum* contigs against each other.
Those contigs that had no match in the other species
(either in contigs or in singletons) were regarded as
potential differentially expressed genes (PDEGs). Starting
bidirectional BLAST search from either species, a similar
number of PDEGs was found: 4245 in *F. esculentum* and
4255 in *F. tataricum*. The PDEG fraction of both tran-
sicomes is enriched in contigs that have no GenBank
matches. Though being quite small (13 genes), this subset
produced well-resolved tree topologies similar to those
inferred in many recent phylogenetic studies. To provide
further validation of this approach and to assess the
category is higher in *F. tataricum* - presumably, because
of higher total number of GO-annotated contigs in this
species. However, one category, “DNA metabolic pro-
cess”, is represented by a considerably higher number of
contigs in *F. esculentum* (79) than *F. tataricum* (49). In
both species, the contigs from this category are almost
exclusively represented by sequences with high similarity
to retrotransposons. Thus there is strong evidence that *F.
esculentum* and *F. tataricum* differ in the number or
expression of retrotransposons.

In addition to the direct comparison of the PDEG sets,
we addressed the question whether they represent random
subsets of the overall transcriptomic datasets. To test this,
the enrichment analysis was performed, with the PDEG
set taken as a test group and the overall set taken as a
reference group (*p* = 0.05). No enrichment by any specific
category was found for *F. tataricum*; while several
categories were found to be overrepresented in the
*F. esculentum* PDEG set (Table 4), in particular categories
obviously related to retrotransposons (DNA integration,
RNA-directed DNA polymerase activity) and to processes
inherent to plants such as sugar metabolism, signaling and
amino acid transport.

When we observe that a certain gene in either species
is sampled by at least one read but its ortholog in the
other species is not sampled does not necessarily
means the absence of expression of this gene in the
other species. It might be not sampled due to technical
issues related to the processing of RNA and cDNA
before sequencing. Taking this into account we per-
formed a statistical assessment of the fraction of such
cases (false positives). Since there is a correlation
between the number of reads constituting contigs of
orthologous genes (see above), the expected number of
false positives differs for different levels of coverage.
Our analysis indicates that only about 25-30% of PDEG
indeed represent genes whose orthologs are not
expressed in the other species. For genes that are repres-
ented by a high number of reads in one species the
probability is much higher (Table 5).

### Shared single-copy genes in the 454 transcriptome assemblies and their phylogenetic analysis

Large-scale transcriptome data are a potential source of
information for multigene phylogenetic analysis (the
phylogenomic approach). To adopt its use in plant phy-
genetics, Duarte et al. [27] identified a set of single-
copy genes shared between *Arabidopsis*, *Populus*, *Vitis*
and *Oryza*. Only few of them were well represented in
EST assemblies across the major lineages of angios-
perms. Thus being quite small (13 genes), this subset
produced well-resolved tree topologies similar to those
inferred in many recent phylogenetic studies. To provide
further validation of this approach and to assess the

### Table 2 Summary of BLAST search and annotation

|                        | Fagopyrum esculentum | Fagopyrum tataricum |
|------------------------|-----------------------|---------------------|
| Contigs                |                       |                     |
| Number of contigs with BLAST match | 19122 | 19072 |
| Number of contigs with GO annotation | 15115 | 14684 |
| Number of contigs without BLAST matches | 6313 | 6329 |
| Number of contigs with predicted ORF/mean length of ORF, bp | 25345/368 (30-1637) | 24984/390 (39-1983) |
| Species provided most of all best BLAST hits | Vitis vinifera | Vitis vinifera |

| Singletons             |                       |                     |
| Total number of singletons | 56766 | 42899 |
| Number of singletons with BLAST match | 14260 | 14671 |
| Number of singletons with GO annotation | 10303 | 10448 |
| Number of singletons without BLAST matches | 42506 | 28228 |
| Species provided most of all best BLAST hits | Vitis vinifera | Vitis vinifera |
Figure 3 Distribution of Gene Ontology categories for Biological Process (a) and Molecular Function subontologies (b) for *F. esculentum* (violet bars) and *F. tataricum* (blue bars).
Table 3 Putative orthologs of buckwheat genes involved in flower development (defined according to Alvarez-Buylla et al. 2008) identified by bidirectional BLAST search

| Arabidopsis gene | number | F. esculentum | F. tataricum |
|------------------|--------|---------------|--------------|
| APETALA1         | AT1G669120 | Fagopyrum_escapedum_MIRA_VER3_c1474 | Fagopyrum_tataricum_MIRA_VER3_c880 |
| APETALA2         | AT4G366920 | Fagopyrum_escapedum_MIRA_VER3_c15879 | Fagopyrum_tataricum_MIRA_VER3_c21103 |
| APETALA3         | AT3G54340 | Fagopyrum_escapedum_MIRA_VER3_c5503 | Fagopyrum_tataricum_MIRA_VER3_c2664 |
| PISTILLATA       | AT5G20240 | Fagopyrum_escapedum_MIRA_VER3_c1334 | Fagopyrum_tataricum_MIRA_VER3_c2417 |
| AGAMOUS          | AT4G18960 | Fagopyrum_escapedum_MIRA_VER3_c3409 | Fagopyrum_tataricum_MIRA_VER3_c2578 |
| SEPALLATA1       | AT5G15800 | Fagopyrum_escapedum_MIRA_VER3_c876 | Fagopyrum_tataricum_MIRA_VER3_c232 |
| SEPALLATA3       | AT1G24260 | Fagopyrum_escapedum_MIRA_VER3_c1493 | Fagopyrum_tataricum_MIRA_VER3_c2041 |
| WUSCHEL          | AT2G17950 | Fagopyrum_escapedum_MIRA_VER3_c13614 | |
| SUPERMAN          | AT3G23130 | Fagopyrum_escapedum_MIRA_VER3_c21331 | Fagopyrum_tataricum_MIRA_VER3_c11424 |
| ARGONAUTE1       | AT1G48410 | Fagopyrum_escapedum_MIRA_VER3_c1224 | Fagopyrum_tataricum_MIRA_VER3_c788 |
| TSO1             | AT3G22780 | Fagopyrum_escapedum_MIRA_VER3_c13853 | |
| CLAVATA1         | AT1G08590 | Fagopyrum_escapedum_MIRA_VER3_c8814 | |
| CLAVATA2         | AT1G65380 | Fagopyrum_escapedum_MIRA_VER3_c20999 | Fagopyrum_tataricum_MIRA_VER3_c16742 |
| FASCIATA1        | AT1G65470 | Fagopyrum_escapedum_MIRA_VER3_c10249 | Fagopyrum_tataricum_MIRA_VER3_c18675 |
| FASCIATA2        | AT5G64630 | Fagopyrum_escapedum_MIRA_VER3_c21840 | |
| SHOOT MERISTEMLESS | AT1G62360 | Fagopyrum_escapedum_MIRA_VER3_c5954 | |
| ULTRAPETALAL1    | AT4G28190 | Fagopyrum_escapedum_MIRA_VER3_c14864 | Fagopyrum_tataricum_MIRA_VER3_c12493 |
| WIGGLY/ENHANCED RESPONSE TO ABSCICIC ACID1 | AT5G40280 | Fagopyrum_escapedum_MIRA_VER3_c7673 | Fagopyrum_tataricum_MIRA_VER3_c5603 |
| LEUNIG           | AT4G32551 | Fagopyrum_escapedum_MIRA_VER3_c11836 | Fagopyrum_tataricum_MIRA_VER3_c4980 |
| LEUNIG HOMOLOG   | AT2G32700 | Fagopyrum_escapedum_MIRA_VER3_c14461 | Fagopyrum_tataricum_MIRA_VER3_c7645 |
| SEUSS            | AT1G43850 | Fagopyrum_escapedum_MIRA_VER3_c3409 | Fagopyrum_tataricum_MIRA_VER3_c2578 |
| YABBY1           | AT2G41190 | Fagopyrum_escapedum_MIRA_VER3_c8803 | |
| AS1              | AT2G37630 | Fagopyrum_escapedum_MIRA_VER3_c2262 | Fagopyrum_tataricum_MIRA_VER3_c2417 |
| AUXIN RESPONSE FACTOR 5/ MONOPTEROS | AT1G19850 | Fagopyrum_escapedum_MIRA_VER3_c12988 | |
| PHABULOSA        | AT2G34710 | Fagopyrum_escapedum_MIRA_VER3_c6862 | |
| YABBY3           | AT4G00180 | Fagopyrum_escapedum_MIRA_VER3_c6676 | Fagopyrum_tataricum_MIRA_VER3_c9095 |
| AINTEGUMENTA     | AT4G37750 | Fagopyrum_escapedum_MIRA_VER3_c21840 | Fagopyrum_tataricum_MIRA_VER3_c22133 |
| ARF3/ETTN        | AT2G33860 | Fagopyrum_escapedum_MIRA_VER3_c22133 | |
| PGP19            | AT3G28860 | Fagopyrum_escapedum_MIRA_VER3_c2262 | Fagopyrum_tataricum_MIRA_VER3_c326 |
| PERIANTHA        | AT1G68640 | Fagopyrum_escapedum_MIRA_VER3_c7084 | |
| PETAL LOSS       | AT5G03680 | Fagopyrum_escapedum_MIRA_VER3_c8866 | Fagopyrum_tataricum_MIRA_VER3_c32716 |
| PIN-FORMED1      | AT1G73590 | Fagopyrum_escapedum_MIRA_VER3_c3355 | |
| PIN-FORMED3      | AT1G70940 | Fagopyrum_escapedum_MIRA_VER3_c10169 | Fagopyrum_tataricum_MIRA_VER3_c4499 |
| PINOID           | AT2G34630 | Fagopyrum_escapedum_MIRA_VER3_c11571 | Fagopyrum_tataricum_MIRA_VER3_c4925 |
| TOUSLED          | AT5G20930 | Fagopyrum_escapedum_MIRA_VER3_c11648 | |
| ABORTED MICROSPORES | AT2G16910 | Fagopyrum_escapedum_MIRA_VER3_c11412 | |
| AUXIN RESPONSE FACTOR 6 | AT1G30330 | Fagopyrum_escapedum_MIRA_VER3_c14779 | Fagopyrum_tataricum_MIRA_VER3_c11412 |
| AUXIN RESPONSE FACTOR 8 | AT5G37020 | Fagopyrum_escapedum_MIRA_VER3_c6633 | Fagopyrum_tataricum_MIRA_VER3_c11597 |
| BARELY ANY MERISTEM 1 | AT5G65700 | Fagopyrum_escapedum_MIRA_VER3_c2493 | Fagopyrum_tataricum_MIRA_VER3_c95 |
| CORONATINE INSENSITIVE 1 | AT2G39940 | Fagopyrum_escapedum_MIRA_VER3_c11571 | Fagopyrum_tataricum_MIRA_VER3_c15577 |
| DELAYED DEHISCENCE1/OPDA-REDUCTASE 3 | AT2G06050 | Fagopyrum_escapedum_MIRA_VER3_c3772 | Fagopyrum_tataricum_MIRA_VER3_c566 |
| FATTY ACID DESATURASE 7 | AT3G11170 | Fagopyrum_escapedum_MIRA_VER3_c1351 | Fagopyrum_tataricum_MIRA_VER3_c705 |
| FAD8             | AT5G05580 | Fagopyrum_escapedum_MIRA_VER3_c705 | |
utility of 454 sequencing data for phylogenomics we performed a search for the orthologs of these genes in five 454 transcriptome assemblies. Two of them are those reported in this study - *F. tataricum* and *F. esculentum*, the third one is from another Caryophyllales species - *Silene latifolia* (sequences provided by D. A. Filatov). Taking into account that the analysis of plastid genes indicates on the affinity of Caryophyllales and asterids [10,28] two recently published transcriptomes of asterids: *Craterostigma plantagineum* [22] and *Artemisia annua* [29] were also included in the analyses. For all species almost all genes were present: 13 in *Artemisia annua*, 12 in *F. tataricum*, *Craterostigma plantagineum* and *Silene latifolia* and 11 in *F. esculentum*. For most genes, only one sequence with high similarity to a certain gene was found thus suggesting that they are single-copy in the species sampled. The only exception is *Artemisia annua*, where each of two genes - the orthologs of AT4G08230 and AT5G63135 - is represented by two sequences with high similarity to the *Arabidopsis* gene but differing one from another by multiple nucleotide substitutions and indels. We assume that these sequences represent paralogs emerged from the recent duplication (presumably, duplication occurred at the level of the genus since no paralogs were found in the transcriptomes of another species from Asteraceae). Such «shallow paralogs» do not necessarily adversely affect phylogenetic reconstruction [30] so they were included in the analysis (under the names of *Artemisia annua1* and *Artemisia annua2*). The sequences were

**Table 3** Putative orthologs of buckwheat genes involved in flower development (defined according to Alvarez-Buylla et al. 2008) identified by bidirectional BLAST search (Continued)

| Ortholog | Buckwheat Accession | Fagopyrum tataricum MIRA_VER3 Accession |
|----------|---------------------|----------------------------------------|
| GA INSENSITIVE DWARF1A | AT3G05120 | Fagopyrum_tataricum_MIRA_VER3_c0120 |
| GA INSENSITIVE DWARF1B | AT3G63010 | Fagopyrum_esculentum_MIRA_VER3_c0789 |
| GA INSENSITIVE DWARF1C | AT5G27320 | Fagopyrum_tataricum_MIRA_VER3_c0120 |
| MALE STERILE1 | AT5G22260 | Fagopyrum_tataricum_MIRA_VER3_c1838 |
| MYB21 | AT5G27810 | Fagopyrum_tataricum_MIRA_VER3_c0806 |
| RECEPTOR-LIKE PROTEIN KINASE 2 | AT3G02130 | Fagopyrum_esculentum_MIRA_VER3_c05605 |
| BREVIPECCELLUS/KNAT1 | AT4G08150 | Fagopyrum_esculentum_MIRA_VER3_c13223 |
| CRABS CLAW | AT1G69180 | Fagopyrum_tataricum_MIRA_VER3_c1676 |
| NGATHA1 | AT2G46870 | Fagopyrum_tataricum_MIRA_VER3_c0303 |
| NO TRANSMITTING TRACT | AT3G7670 | Fagopyrum_esculentum_MIRA_VER3_c16587 |
| BLR | AT5G20230 | Fagopyrum_tataricum_MIRA_VER3_c2133 |
| STK | AT4G09960 | Fagopyrum_tataricum_MIRA_VER3_c6138 |
| Dicer-LIKE1 | AT1G01040 | Fagopyrum_tataricum_MIRA_VER3_c24560 |
| SPATULA | AT4G36930 | Fagopyrum_esculentum_MIRA_VER3_c7255 |
| STY1 | AT2G1060 | Fagopyrum_tataricum_MIRA_VER3_c19680 |
| STY2 | AT4G36260 | Fagopyrum_tataricum_MIRA_VER3_c6938 |
| BIG BROTHER | AT3G63530 | Fagopyrum_esculentum_MIRA_VER3_c12106 |
| HWS | AT3G61590 | Fagopyrum_esculentum_MIRA_VER3_c12481 |

**Table 4** Characteristics of potential differentially expressed genes (PDEG) fraction of buckwheat transcriptomes

| Transcriptome | Fagopyrum esculentum | Fagopyrum tataricum |
|---------------|----------------------|---------------------|
| total number of PDEG | 4245 | 4255 |
| without BLAST matches | 2567 | 2037 |
| with GO annotation | 1132 | 1588 |
| number of contigs covered by more than 10 reads | 158 | 52 |
added to the dataset from [27]. The resulting 73-taxon nucleic acid sequences dataset contained 7869 characters (including gaps). The phylogeny was reconstructed using the maximum parsimony (MP) and Bayesian inference (BI) methods based on both nucleotide and amino acid sequences. As an alternative to the combined approach a super distance matrix (SDM) analysis was also applied (only to nucleotide sequences). The trees resulting from MP and BI were similar to those inferred by Duarte and coworkers [27] in topology and resolution (Figure 5). The maximum parsimony trees inferred from both nucleotide and amino acid sequence data were much less resolved, and much less consistent with the current knowledge on the angiosperm phylogeny. The Bayesian trees were well resolved, with posterior probabilities higher than 0.5 for all nodes in the amino acid tree and all but one in the nucleotide tree. In general, the Bayesian tree is similar to those inferred from the analysis of several genes and from the multigene analysis of plastid datasets [28], with Amborella trichopoda and Nuphar advena being basal among angiosperms, and monocots, eurosids and asterids resolved as monophyletic. Nonetheless there are some points of

### Table 5 Expected number of differentially expressed genes in the PDEG sets

| Fagopyrum esculentum | observed number of PDEG | expected number of differentially expressed genes |
|----------------------|-------------------------|-----------------------------------------------|
| 2                    | 2311                    | 524.672773                                   |
| 3                    | 788                     | 393.71303                                    |
| 4                    | 342                     | 242.460795                                   |
| 5                    | 207                     | 178.845566                                   |
| 6                    | 134                     | 121.243316                                   |
| 7                    | 82                      | 78.293171                                    |
| 8                    | 61                      | 59.094585                                    |
| > 8                  | 320                     | 320                                           |

| Fagopyrum tataricum |
|---------------------|
| 2                   | 2456                  |
| 3                   | 838                   |
| 4                   | 406                   |
| 5                   | 183                   |
| 6                   | 110                   |
| > 6                 | 262                   |

Figure 5 Trees inferred from the Bayesian analysis (a) and from the maximum parsimony analysis (b) of nucleotide sequences of 13 single-copy genes. Branch lengths are proportional to the number of expected nucleotide substitutions; scale bar corresponds to one substitution per ten sites for the Bayesian tree and to 100 changes for maximum parsimony tree. Posterior probability and bootstrap values (greater than 0.5 and 50% respectively) are indicated by the colored bars placed on branches.
incongruence, primarily the position of magnoliids and non-monophyly of eurosids I and II.

An interesting result is the position of the Caryophyllales. With the addition of *Fagopyrum* and *Silene*, the sister relationships of the Caryophyllales gained high support (PP = 0.96 for nucleotide data and 1 for amino acid) in the Bayesian analysis. This is consistent with our previous data on the plastid genome phylogeny [10], but in the nuclear gene analysis this topology is present as highly supported for the first time. In the dataset of Duarte et al. [27], where the only caryophyllid species, *Beta vulgaris* was present, it was also grouped with asterids in the maximum likelihood tree but with very a low support (< 50%).

The supertree resulting from SDM is in most points congruent with the Bayesian trees and also supports sister relationships of the Caryophyllales and asterids (Additional file 5).

**Discussion**

**Utility of 454 sequencing for gene discovery in non-model species**

Despite the economical importance of buckwheat and the great advances in DNA sequencing technologies no genomic data were available to date for any of the species of *Fagopyrum*, with the exception of the complete chloroplast genome sequence of *F. esculentum* ssp. *ancestrale* [10]. Prior to this study, the total number of sequences in GenBank was 149 for *F. esculentum* and 121 (including 83 unannotated EST) for *F. tataricum*. Sequencing and assembly of genes expressed in flower and inflorescence presented in this study results in 25 thousand contigs for each species. Given that the average number of genes encoded in a plant nuclear genome is about 30 thousands (as estimated from seven completely sequenced genomes), we assume that our dataset represents a substantial fraction of the buckwheat genes. Among them are putative orthologs of genes that play key roles in the flower and inflorescence development, including those that are expressed at low level and are not widely represented in other transcriptome assembles. These results are highly consistent with previous experimental and simulation-based studies [31-33] and provide further support for the use of combination of cDNA normalization and 454 sequencing for fast transcriptome characterization in non-model species. Application of the floral transcriptome sequence data for gene expression studies in *F. esculentum* mutants with altered flower development [34-36] will constitute the basis for the study of the genetic control of flower development in this species.

A fraction of contigs that did not produce any BLAST hits may also be an integral part of genomic data. The existence of such “non-blastable” sequences is reported in virtually all plant transcriptomes characterized to date, with their fraction varying from about 30 to 70% [23-25], depending on the species, the depth of sequencing and the parameters of BLAST search. There are two major groups of effects that may be responsible for this. The first group comprises the technical issues such as low quality of raw data, contamination by genomic DNA, inaccurate contig assembly and wrong choice of search parameters (e.g. too stringent e-value cut-off). The second group of effects, biological ones, is related to the characteristics of the data that are inherent to the species: the existence of non-coding cDNA fraction (primarily 5’ and 3’UTR), lineage-specific genes (i.e. genes that are present in the genome being considered but absent from genomes represented in the databases) and fast-evolving genes, those that have orthologs in the other species but with a high level of sequence divergence precluding efficient recognition of orthologs.

In our study we tried to minimize the influence of technical effects. DNAse treatment of RNA was performed before cDNA synthesis to preclude genomic DNA contamination; also low quality data were excluded from the assembly process. As for the assembly accuracy, in the absence of a reference genome the possibility of assembly errors can not be completely ruled out. However given the relatively large length of 454-generated reads even chimeric combination of reads corresponding to different genes in one contig is unlikely to hamper efficient BLAST search. Indeed, assembly errors cause problems with similarity search only if short reads corresponding to small and weakly conserved fragments of different genes are combined into one contig.

Thus we assume that most of non-blastable contigs in our data are due to biological reasons. Some of them might indeed represent non-translated regions, but since the average length of predicted ORF in such contigs is significantly higher than in artificial sequences produced by random shuffling, we suggest that most of them correspond to protein-coding sequences, either lineage specific or highly variable. Currently there are no means to select one of these two options. Addition of more genomic and/or transcriptomic data for the Polygonaceae and other Caryophyllales is expected to improve the annotation and gene prediction in buckwheat.

The favourable effect of using the data from closely related species for annotation was reported for the analysis of chestnut (*Castanea*; Fagaceae) transcriptomes [32]. Several genes that could not be annotated using the information from the *Arabidopsis* proteome (*Castanea* and *Arabidopsis* belong to different clades of rosids - eurosids I and eurosids II) were successfully annotated when compared with *Populus* (belonging to eurosids I) proteome. This emphasizes that despite the availability of large-scale sequence data for model plant species and the increased rate of data acquisition for non-model
species denser taxon sampling will further improve plant gene annotation. Given that the Caryophyllales is a group of exceptional interest from the point of view of evolutionary developmental biology we expect that more transcriptomic data will be available for this group.

Comparison of two buckwheat transcriptomes: finding potential differentially expressed genes

Gene expression is often compared within one species at different developmental stages or in different conditions [22,37], and several efficient tools have been developed to analyze the enrichment in specific sets of genes (reviewed in [38]). The interspecific comparisons are rare and confined to the cases when it is possible to use sequence information from a phylogenetically close model species [39,40]. The possibility to use de novo sequencing data for the comparison of gene expression (even on the qualitative level, that is, to identify the presence or absence of expression) in different species has, to our knowledge, never been addressed in plant genomics.

Here we used an approach that is similar in its concept to the digital transcriptome subtraction [41]. The latter is based on the similarity search of sequences from one species against the other and selection of sequences that do not have significant matches. As expected, for most sequences from both datasets the sequences with high similarity were found in another species, either in contigs, or in singletons. However, about four thousands contigs in each species were found to be “species-specific”. It is difficult to directly associate these contigs with the observed differences between *F. esculentum* and *F. tataricum*, the more so because about two thirds of them do not have significant BLAST hits in the nr database, and thus lack any functional annotation.

Still, some patterns are likely caused by species-specific expression. Firstly, the *F. esculentum* PDEG set shows a strong prevalence of contigs highly similar to retrotransposons. This suggests that either the *F. esculentum* genome has more retrotransposons than the genome of *F. tataricum*, or more of them are expressed. The former explanation seems to be more plausible, considering the fact that the genome size of *F. esculentum* is three-fold larger than that of *F. tataricum* [42]. While such differences in the genome size in plants are often caused by the polyploidy, it is not the case for the buckwheat since both species have the same chromosome number [42,43].

Secondly, the *F. esculentum* PDEG set is enriched in genes related to the disaccharide metabolism. Mono- and disaccharides, primarily glucose, fructose and sucrose, are a major component of the *F. esculentum* nectar [44]. *F. tataricum*, being a self-pollinated species, does not produce nectar. So the disaccharide metabolism genes indeed might be differentially expressed due to the difference in the nectar production. However, any conclusion about the differential expression should be treated with caution and verified by more precise methods such as quantitative RT-PCR. The set of “unique” contigs identified in our study is a mixture of genes whose orthologs are actually not expressed in the other species and those genes that are expressed in both species but are absent in the sequence data from one of the species due to some random fluctuations on different stages of RNA and cDNA processing. Indeed, our statistical analysis indicates that about 25-30% of PDEG are really differentially expressed. One of the possible issues responsible for this is the normalization because in normalized libraries more transcripts are represented and thus the probability of being sequenced is lower for each transcript than in case of non-normalized library.

The probability that a gene is expressed, given the observations of no reads from the other genome, is proportional to its coverage. It is very low for low-covered genes but among genes represented by a contig covered by more than 7 at least 95% of genes are expected to be differentially expressed (Table 5). Thus, further investigation of the genetic basis of specific differences between *F. esculentum* and *F. tataricum* may be conducted in two ways. The first one is the survey of the expression of PDEGs from the categories that are enriched in either species and that are likely to have species-specific expression for biological reasons (retrotransposons, sugar metabolism genes). The second one is the analysis of those genes that have a high number of reads in one transcriptome but are absent in the other. A combination of these approaches may provide insights in the genetic and evolutionary mechanisms underlying morphological and physiological differentiation in the genus *Fagopyrum*. Further insights, including quantitative analysis of gene expression levels are also expected to result from the sequencing of non-normalized cDNA. This approach demonstrated its utility on several plant species [37,45].

Phylogenetic utility of transcriptome sequence data

Besides its potential utility for functional genetic studies, the data from large-scale transcriptome sequencing is an important source of information for phylogenetic analysis. Phylogenetic studies based on transcriptome sequences yielded well-resolved and highly supported tree topologies for many groups of animals [46]. In plants, however, the phylogenomic approach has been mostly limited to plastid genome sequences [28]. There are two reasons for that. The first one is the relative scarcity of plant genome-scale data. Secondly, the starting point of any phylogenetic analysis, identification of orthologous genes, is a major problem for plants due to abundant genome-wide and small scale duplications.
Recently a set of single-copy genes was proposed for phylogenetic analysis on different taxonomic level. Though a number of articles reporting de novo plant transcriptome sequence data were published [25,33] none of them reported the use of these data for phylogenetic analysis. Here we attempted to validate the use of these genes for inferring angiosperm phylogeny by including in the analysis the data from four 454 transcriptome assemblies, with focus on Caryophyllales and asterids. Orthologs of most of these genes were found in all these transcriptomes. The tree inferred from the Bayesian analysis of these genes is similar to the ML tree from Duarte et al. [27] but has a higher resolution and node support. We suggest the increased resolution and node support reflects favourable effect of improved taxon sampling.

A major problem observed in our phylogenetic trees is the incongruence between different methods. The Bayesian inference and supertree (SDM) approach yield congruent topologies but these topologies are not congruent with those inferred from the maximum parsimony analysis. Also, MP analysis results in a tree with very low bootstrap support values. This may indicate that the MP is not applicable for transcriptomic data, where high variation is combined with lot of missing data. It is known that several types of data may require additional adjustments of procedures used for phylogenetic analysis (primarily alignment). For example, for non-coding sequences of plastid genomes special rules of alignment were developed and this increased resolution and reliability of phylogenetic trees inferred from these sequences [47,48]. The evolutionary peculiarities of plastid sequences are well studied [47,49]. On the contrary, little is known about the nuclear genes used for the phylogenetic analysis in this study. It is probable that the angiosperm-wide survey of the patterns of evolution of these genes will identify the source of incongruence between different methods of the analysis and, as the result, improve their utility for plant phylogenetics.

However, even under limited taxon sampling these genes provide valuable information on phylogenetic relationships within the angiosperms. The novel result inferred from these data is the strong support of sister relationships of Caryophyllales (including buckwheat) and asterids. Earlier we have shown that these relationships are strongly supported by the chloroplast genome sequence data [10]. It is well known however that the results of phylogenetic analysis of chloroplast gene datasets may be misleading and thus require verification by independent data [50,51]; the present study reports such verification for the phylogenetic position of Caryophyllales.

The results presented here emphasize that even shallow-coverage transcriptome sequence data are an important source of information for phylogenetic applications.

The addition of such data from various lineages of flowering plants is expected to improve greatly the resolution, support and reliability of phylogenetic trees and to provide novel insights into the evolution of angiosperms.

Conclusions

454 pyrosequencing of normalized cDNA libraries produced a large set of cDNA sequences for two congeneric plant species, Fagopyrum esculentum and F. tataricum. These sequences are an important resource for the evolutionary and developmental genetics in these species and in the order Caryophyllales. Analysis of single copy genes derived from transcriptome sequence data have great potential for inferring angiosperm phylogeny, especially with increased taxonomic sampling.

Methods

RNA extraction

Total RNA was extracted from developing inflorescences of Fagopyrum esculentum ssp. ancestrale and F. tataricum using the RNEasy Plant Mini kit (Qiagen) with following modifications: 50 μl of Plant RNA Isolation Aid reagent (Ambion) and 10 μl of b-mercaptoethanol were added to the lysis buffer RLT prior to homogenization and the homogenization was performed without liquid nitrogen. All other steps were performed according to the manufacturer’s instructions. To avoid genomic DNA contamination RNA was treated with RNase-free DNase I (Qiagen). The RNA integrity was assessed by the agarose gel electrophoresis (1% gel with addition of SYBRGreen dye).

cDNA synthesis, amplification and normalization and 454 sequencing

Total RNA was used for double-stranded cDNA synthesis using the SMART approach [52]. To increase the representation of weakly expressed genes, SMART-prepared amplified cDNA was then normalized using the DSN normalization method [53]. Normalization included cDNA denaturation/reassociation, treatment by a duplex-specific nuclease (DSN, [54]) and amplification of the normalized fraction. cDNA synthesis and normalization were performed by the Evrogen company, the detailed protocols are available in Additional file 6.

The efficiency of normalization was tested by real-time PCR. Fragments of GAPDH (a housekeeping gene expressed at a high level) and of the WUS ortholog (encodes a transcription factor expressed at a low level) were amplified from the normalized and non-normalized libraries. The 32-fold decrease of GAPDH expression was observed in the normalized library. Then 7 μg of normalized cDNA were fractionated and sequenced using 454 GS-FLX sequencer with 454 Titanium chemistry. Sequencing was performed by The Centre for
Applied Genomics, The Hospital for Sick Children, Toronto, Canada. Each sample was run on a half of a picotiter plate.

Assembly, annotation and comparative analysis of transcriptomes

SeqClean [55] was used to remove polyA sequences from the raw data. We did not use SeqClean to remove adapter sequences since it works only with 100-92% identity segments and in our case adapters contained many sequencing errors caused by the homopolymer runs. To remove the adapters with less than 92% identity we developed two scripts, which are available upon request. MIRA assembler [56] version 3.0.0rc4 was used for the 454 data assembly. The assembler was run in its ‘accurate’ mode with the assembly type set as ‘EST’.

The resulting contigs were subject to the BLASTX search against the non-redundant protein database (nr) with the e-value threshold of $10^{-6}$ and the HSP length cut-off of 50, as implemented in the BLAST2Go program [57]. The ORF Predictor tool was used to identify ORFs (http://proteomics.ysu.edu/tools/OrfPredictor.html, [58]). Searches against the Arabidopsis proteome (e-value threshold of $10^{-6}$) and the searches between F. esculentum and F. tataricum (e-value of $10^{-15}$) were performed using BLASTX/TBLASTN and BLASTN, respectively, as implemented in BioEdit 7.0.9 [59]. The Gene Ontology annotation [24] was done using BLAST2Go with the annotation cut-off of $10^{-6}$. We used simplified GOslim (plant) annotations (developed by TAIR, [60]) to select genes expressed in the flower and in the inflorescence. For the identification of genes that are represented in the transcriptome of one species of buckwheat but are not represented in the other (here and further called potential differentially expressed genes, PDEG), a Perl script was developed that processes the BLAST output and extracts sequences that do not have significant matches. The enrichment analysis was performed using Fisher’s exact test as implemented in Gossip/BLAST2Go [61]. The statistical assessment of the number of false positives in the PDEG sets was performed as follows. First, the probability $p(i,j,s)$ to observe $i$ reads for the gene in one species ($s$) given $j$ reads for its ortholog in the other species was calculated. For each possible $j$ and $s$, the observed number of orthologous pairs with $j$ reads for gene in species $s$ fitted well the Poisson model after filtering out a small number of outliers. The model parameters $\lambda(i,j,s)$ were estimated from these data using the maximum-likelihood method, taking into account that pairs with zero number of reads for any of the genes were not observed for obvious reasons. Finally, the number of false positives was calculated as $\sum k(j,s) p(0,j,s)$ where $k(j,s)$ is the total number of genes with $j$ reads in species $s$.

Phylogenetic analysis

For the phylogenetic analysis, a set of 13 single copy genes from 69 taxa, from the [27] study was used. The orthologs of these genes were found by the BLAST search of corresponding Arabidopsis genes in three 454-sequenced transcriptomic datasets of Caryophyllales, F. esculentum, F. tataricum, and Silene latifolia (Filatov D. A., personal communication) and two transcriptomes of asterids: Artemisia annua [29] and Craterostigma plantagineum [22]. Since the orthologs from the two buckwheat species are highly similar (> 95%) the sequences from only one species (F. tataricum) used. Sequences of F. tataricum, S. latifolia, A. annua and C. plantagineum were added to the alignment from [27] then the whole set was checked for frameshifts and the latter, if present, were corrected by inserting N to recover in frame translation. Sequences of transcripts were translated into amino acid sequences by BioEdit and further aligned using MUSCLE ver. 3.6 [62]. The nucleotide sequence alignment was overlaid on the amino acid sequence alignment. Gap-rich positions and positions of ambiguous alignment were excluded from analyses (789 in total). The phylogenetic analysis using the maximum parsimony (MP) method was performed using PAUP* ver. 4.0b8 [63]. The MP analysis involved a heuristic search using TBR branch swapping and 500 random addition replicates. The non-parametric bootstrap analysis [64] was performed with 1000 replicates with TBR branch swapping. The Bayesian inference of phylogeny was explored using the MrBayes program ver. 3.1.2 [65]. The Bayesian analysis was applied both to the amino acid and nucleotide data with two runs with three chains in each. Each chain started with a random tree, 1,000,000 replicates for amino acid data and 5,000,000 replicates for nucleotide data were generated, trees were sampled every 100 generations. The number of discarded trees was determined using the cold-chain log-likelihood examination. The JTT model of amino acid substitutions was determined by the Bayesian information criterion in Modelgenerator ver. 0.43 [66], the GTR+I+Γ model of nucleotide substitution was determined by the Akaike information criterion in Modeltest ver. 3.7 [67], the data were treated as a single partition. As an alternative to the combined analysis a super distance matrix (SDM) method [68] was applied. The SDM-algorithm combines the evolutionary distances obtained from each gene into a single distance supermatrix, therefore for each gene a corresponding model of evolution was determined in Modeltest, then the maximum likelihood distances according to the selected model were computed with PAUP* and the latter was used as an alternative model in the phylogenetic analysis. A super-tree was constructed using Fitch program from PHYLIP package [69].
Additional material

Additional file 1: Fagopyrum esculentum transcriptome sequence data. The sequences of F. esculentum contigs and singletons and the summary of BLAST search and GO annotation of these sequences.

Additional file 2: Fagopyrum tataricum transcriptome sequence data. The sequences of F. tataricum contigs and singletons and the summary of BLAST search and GO annotation of these sequences.

Additional file 3: Coverage of previously identified Fagopyrum esculentum genes. Accession numbers of the sequences of F. esculentum genes known from previous studies and their coverage in the 454 transcriptome assembly.

Additional file 4: Distribution of GO categories in PDEG. Distribution of GO categories for Biological Process (A) and Molecular Function subontologies (B) for F. esculentum (violet bars) and F. tataricum (blue bars) PDEG.

Additional file 5: SDM phylogenetic tree. Phylogenetic tree inferred from the SDM analysis of 13 single-copy nuclear genes in 73 seed plant taxa. Scale bar corresponds to one substitution per ten sites.

Additional file 6: CDNA synthesis, amplification and normalization. Detailed protocol of cDNA synthesis, amplification and normalization.

List of abbreviations
AFLP: amplified fragment length polymorphism; BL: Bayesian inference; DSN: duplex-specific nuclease; GO: Gene Ontology; ML: maximum likelihood; MP: maximum parsimony; PCR: polymerase chain reaction; PDEG: potential differentially expressed genes; PP: posterior probability; RAPD: random amplified polymorphic DNA; RT-PCR: reverse transcription polymerase chain reaction; SDM: super distance matrix; SSR: simple sequence repeat

Acknowledgements
The authors are grateful to Dmitry Shagin (Evrogen Lab, Moscow, Russia) for unpublished data on Silene latifolia gene set used for phylogenetic analysis and to two anonymous reviewers for their helpful comments. The study is supported by the Russian Foundation for Basic Research (project N09-04-01383), the Ministry of Education and Science of Russian Federation (contracts P513 and P515 and project “Phylogenetic analysis of complex selection in molecular evolution”), the President’s program for support of young scientists and the Russian Academy of Science Presidium Program "Molecular and Cell Biology".

Author details
1Department of Evolutionary Biochemistry, A.N. Belozersky Institute of Physico-Chemical Biology, M.V. Lomonosov Moscow State University, Moscow, Russia. 2Evolutionary Genomics Laboratory, Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia. 3V.A. Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia. 4A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Science, Moscow, Russia. 5Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia.

Authors’ contributions
MVD conceived and coordinated the study, performed transcriptome annotation and comparison and drafted the manuscript, ASK assembled contigs and developed scripts for the extraction of PDEG, DWV carried out statistical analysis of the PDEG, THS performed phylogenetic analysis, VJM and MSG participated in data analysis and interpretation and revised the manuscript, AAP participated in the design of the study and interpretation of the results and prepared RNA samples. All the authors read and approved the manuscript.

Received: 18 October 2010 Accepted: 13 January 2011
Published: 13 January 2011

References
1. Ohnishi O: Discovery of the wild ancestor of common buckwheat. Fagopyrum 1991, 11:5-10.
2. Ohnishi O: Search for the wild ancestor of buckwheat. I. Description of new Fagopyrum (Polygonaceae) species and their distribution in China and Himalayan hills. Fagopyrum 1998, 15:18-25.
3. Ohnishi O, Matsuoka Y: Search for the wild ancestor of buckwheat. II. Taxonomy of Fagopyrum (Polygonaceae) species based on morphology, isozymes and cpDNA variability. Genes and Genetic Systems 1996, 71:383-390.
4. Yaoa Y, Wang Y, Ohnishi O, Campbell CG: Amplified fragment length polymorphism linkage analysis of common buckwheat (Fagopyrum esculentum) and its self wild-pollinated relative Fagopyrum homotropicum. Genome 2004, 47:345-351.
5. Sharma R, Jana S: Species relationships in Fagopyrum revealed by PCR-based DNA fingerprinting. Theoretical and Applied Genetics 2002, 105:306-312.
6. Licen M, Krítf J: Buckwheat (Fagopyrum esculentum Moench) low molecular weight seed proteins are restricted to the embryo and are not detectable in the endosperm. Plant physiology and biochemistry 2005, 43:862-865.
7. Yaoa Y, Ohnishi O: Phylogenetic relationships among Fagopyrum species revealed by the nucleotide sequences of the ITS region of the nuclear RNA gene. Genes and Genetic Systems 1998, 7:201-210.
8. Ohnishi O, Matsuoka Y: Intra- and interspecific phylogeny of wild Fagopyrum (Polygonaceae) species based on nucleotide sequences of noncoding regions in chloroplast DNA. American Journal of Botany 2000, 87:573-582.
9. Nishimoto Y, Ohnishi O, Hasegawa M: Topological incongruence between nuclear and chloroplast DNA trees suggesting hybridization in the uruphylle group of the genus Fagopyrum (Polygonaceae). Genes and Genetic Systems 2003, 78:139-153.
10. Logacheva MD, Sarmigullin TH, Dhingra A, Penin AA: Comparative chloroplast genomics and phylogenetics of Fagopyrum esculentum ssp. ancestrale - a wild ancestor of cultivated buckwheat. BMC Plant Biology 2008, 8:59.
11. Ma H, dePamphilis CW: The ABCs of floral evolution. Cell 2000, 101:5-8.
12. Baum DA, Doebely J, Irish VF, Kramer EM: Response: Missing links: the genetic architecture of flower and floral diversification. Trends in Plant Science 2002, 7:31-34.
13. Nutt P, Ziermann J, Hintz M, Neuffer B, Theißen G: mesophyll as a model system to study the evolutionary relevance of floral homoeotic mutants. Plant Systematics and Evolution 2006, 259:217-235.
14. Albert VA, Solts DE, Carlson JE, Farmerie WG, Wall PK, Blat DC, Solow TM, Mueller LA, Landheir LL, Hu Y, Buzgo M, Kim S, Yoo AU, Frohlich MW, Ped-Trevor E, Schlarbaum SE, Bliss BJ, Zhang X, Tankersley SD, Oppenheimer DG, Solts PS, Ma H, dePamphilis CW, Leebens-Mack J: Floral gene resources from basal angiosperms for comparative genomics research. BMC Plant Biology 2005, 5:5.
15. Samimi C, Björkman T, Sintuna D, Blanchard L: Overcoming the barrier to interspecific hybridization of Fagopyrum esculentum with Fagopyrum tataricum. Euphytica 1996, 91:323-330.
16. Brockington SF, Roeke A, Ramdial J, Moore MJ, Crawley S, Dhingra A, Hila K, Solts DE, Solts PS: Phylogeny of the Caryophyllales sensu lato: revisiting hypotheses on pollination biology and perianth differentiation in the core Caryophyllales. International Journal of Plant Science 2009, 170:627-643.
17. DFCI ice_plant (Mesembryanthemum crystallinum) Gene Index. [http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=ice_plant].
18. Silene vulgaris Transcriptome Project. [http://faculty.virginia.edu/silenevulgaris/].
19. Jaillon O, Aurý JM, Noel B, Polcriti A, Clepet C, Casagrande A, Chioune N, Aubourg S, Vitulo N, Rubin C, Vezzi A, Lepage F, Hugueney P, Dalsiva C, Homer D, Mica E, Joubert D, Samimy C, Björkman T, Siritunga D, Blanchard L: The ABCs of floral evolution. Plant Systematics and Evolution 2006, 259:217-235.
Alaux M, Di Gaspero G, Dumais V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pepele G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecheyn A, Scarpelli C, Artiguenave F, Pê ME, Valle G, Morgante M, Cabcio M, Adam-Blondet AF, Weissenbach J, Quétier F, Wincker P: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 2007, 449:463-467.

20. Tuskan GA, Difazio S, Jansson S, Bohmian J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Stierl A, Aerts A, Bhlerano RR, Bhlerano RP, Blaudz D, Boerjan W, Brn A, Brunner A, Busov V, Campbell M, Carlson J, Chaten MP, Chen JG, Cooper D, Coutinho PM, Coudert J, Covet S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, Depamphilis C, Dettler J, Dirks B, Dubchak I, Duplessis S, Eihling J, Ellis B, Gendler K, Goodstein D, Gribkov M, Grimwood J, Groover A, Gunter L, Hambarger B, Heinez B, Helariutta Y, Hennisat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelker C, Kirkpatrick R, Kist M, Kohler A, Kullu U, Larimer F, Leebens-Mack J, Leple JC, Locascio P, Loe Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe P, Pilate G, Poliakov A, Rozasmukija J, Richardson P, Rinaldi C, Ritald K, Rouze P, Ryabyov D, Schmutz S, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terley T, Tsai CJ, Ubertacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yane G, Yn T, Douglas C, Marra M, Sandberg G, Peer Y, Rohle S, Rokhsar D: The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 2006, 313:1596-1604.

21. Ricinus communis genome project. [http://sgc.jcvi.org/projects/msc/rncinus_commons/]

22. Rodriguez MC, Edgard D, Safras HS, Alquezar D, Rasmussen M, Gilbert T, Nelson RH, Bartels D, Mundy J: Transcriptome of the desiccation tolerant resurrection plant Craterostigma plantagineum. Plant Journal 2010, 63:212-228.

23. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA: Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proceeding of the National Academy of Sciences USA 2007, 104:15832-15837.

24. Duarte JM, Wall PK, Edsgard D, Sarfraz HS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH, Bartels D, Mundy J: Transcriptome of the desiccation tolerant resurrection plant Craterostigma plantagineum. Plant Journal 2010, 63:212-228.

25. Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S: De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 2010, 11:262.

26. Ainsworth C, Thangavelu M, Crossley S, Buchanan-Wollaston V, Parker J: Male and female flowers from the dioecious plant Rumex acetosa follow different patterns of MADS-box gene expression. Plant Cell 1995, 7:1383-1395.

27. Duarte JM, Wall PK, Edsgard D, Sarfraz HS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH, Bartels D, Mundy J: Transcriptome of the desiccation tolerant resurrection plant Craterostigma plantagineum. Plant Journal 2010, 63:212-228.

28. Ainsworth C, Thangavelu M, Crossley S, Buchanan-Wollaston V, Parker J: Male and female flowers from the dioecious plant Rumex acetosa follow different patterns of MADS-box gene expression. Plant Cell 1995, 7:1383-1395.

29. Duarte JM, Wall PK, Edsgard D, Sarfraz HS, Alquezar D, Rasmussen M, Gilbert T, Nielsen BH, Bartels D, Mundy J: Transcriptome of the desiccation tolerant resurrection plant Craterostigma plantagineum. Plant Journal 2010, 63:212-228.

30. Bailey CD, Carr TG, Harris SA, Hughes CE: Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proceedings of the National Academy of Sciences USA 2010, 107:4623-4628.

31. Ainsworth C, Thangavelu M, Crossley S, Buchanan-Wollaston V, Parker J: Male and female flowers from the dioecious plant Rumex acetosa follow different patterns of MADS-box gene expression. Plant Cell 1995, 7:1383-1395.

32. Ainsworth C, Thangavelu M, Crossley S, Buchanan-Wollaston V, Parker J: Male and female flowers from the dioecious plant Rumex acetosa follow different patterns of MADS-box gene expression. Plant Cell 1995, 7:1383-1395.

33. Ainsworth C, Thangavelu M, Crossley S, Buchanan-Wollaston V, Parker J: Male and female flowers from the dioecious plant Rumex acetosa follow different patterns of MADS-box gene expression. Plant Cell 1995, 7:1383-1395.
Shagin DA. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Research 2004, 32:e37.

Shagin DA, Rebrikov DV, Kozhemyako VB, Altshuler IM, Shcheglov AS, Zhulidov PA, Bogdanova EA, Staroverov DB, Rasokazov VA, Lukyanov S: A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. Genome Research 2002, 12:1935-1942.

DFCI Gene Indices Software Tools. [http://compbio.dfci.harvard.edu/cgi/software/].

Chevreux B, Wetter T, Suhai S: Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics 1999, 99:45-56.

Conesa A, Gótz S, García-Gómez JM, Terol J, Talón M, Robles M: Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005, 21:3674-3676.

Min XJ, Butler G, Storms R, Tsang A: OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Research 2005, W677-W680.

DFCI Gene Indices Software Tools. [http://compbio.dfci.harvard.edu/cgi/software/].

Chevreux B, Wetter T, Suhai S: Genome sequence assembly using trace signals and additional sequence information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics 1999, 99:45-56.

Conesa A, Gótz S, García-Gómez JM, Terol J, Talón M, Robles M: Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005, 21:3674-3676.

Min XJ, Butler G, Storms R, Tsang A: OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Research 2005, W677-W680.

61. Blüthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D: Biological profiling of gene groups utilizing Gene Ontology. Genome Informatics 2005, 16(1):106-15.

62. Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 2004, 32:1792-1797.

63. Swofford DL: PAUP*: Phylogenetic analysis using parsimony (*and other methods), ver. 4.0. Sunderland MA: Sinauer Associates; 2003.

64. Felsenstein J: Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985, 39:783-791.

65. Ronquist F, Huelsenbeck JP: MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003, 19:1572-1574.

66. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mchernes JO: Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evolutionary Biology 2006, 6:29.

67. Posada D, Crandall KA: MODELTEST: testing the model of DNA substitution. Bioinformatics 1998, 14:817-818.

68. Criscuolo A, Berry V, Douzery EJ, Gascuel O: SDM: a fast distance-based approach for (super) tree building in phylogenomics. Systematic Biology 2006, 55:740-755.

69. Felsenstein J: PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle; 2004.

doi:10.1186/1471-2164-12-30

Cite this article as: Logacheva et al: De novo sequencing and characterization of floral transcriptome in two species of buckwheat (Fagopyrum). BMC Genomics 2011 12:30.

Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit