

## The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry)

--Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-17-00155R3
<b>Full Title:</b>	The genome sequence and transcriptome of <i>Potentilla micrantha</i> and their comparison to <i>Fragaria vesca</i> (the woodland strawberry)
<b>Article Type:</b>	Data Note
<b>Funding Information:</b>	
<b>Abstract:</b>	<p>The genus <i>Potentilla</i> is closely related to that of <i>Fragaria</i>, the economically important strawberry genus. <i>Potentilla micrantha</i> is a species that does not develop berries, but shares numerous morphological and ecological characteristics with <i>F. vesca</i>. These similarities make <i>P. micrantha</i> an attractive choice for comparative genomics studies with <i>F. vesca</i>. In this study, the <i>Potentilla micrantha</i> genome was sequenced and annotated, and RNA-Seq data from the different developmental stages of flowering and fruiting were used to develop a set of gene predictions. A 327 Mbp sequence and annotation of the genome of <i>P. micrantha</i>, spanning 2,674 sequence contigs, with an N50 size of 335,712, estimated to cover 80% of the total genome size of the species was developed. The genus <i>Potentilla</i> has a characteristically larger genome size than <i>Fragaria</i>, but the recovered sequence scaffolds were remarkably collinear at the micro-syntenic level with the genome of <i>F. vesca</i>, its closest sequenced relative. A total of 33,602 genes were predicted, and 95.1% of BUSCO genes were complete within the presented sequence. Thus, we argue that the majority, of the gene-rich regions of the genome have been sequenced. Comparisons of RNA-Seq data from the stages of floral and fruit development revealed genes differentially expressed between <i>P. micrantha</i> and <i>F. vesca</i>. The data presented are a valuable resource for future studies of berry development in <i>Fragaria</i> and the Rosaceae and they also shed light on the evolution of genome size and organization in this family.</p>
<b>Corresponding Author:</b>	Daniel James Sargent, PhD  UNITED KINGDOM
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Matteo Buti, PhD
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	<p>Matteo Buti, PhD</p> <p>Marco Moretto, PhD</p> <p>Elena Barghini, PhD</p> <p>Flavia Mascagni, PhD</p> <p>Lucia Natali, PhD</p> <p>Matteo Brilli, PhD</p> <p>Alexandre Lomsadze, PhD</p> <p>Paolo Sonogo, PhD</p> <p>Lara Giongo, PhD</p> <p>Michael Alonge, MSc</p>

	Riccardo Velasco, PhD
	Claudio Varotto, PhD
	Nada Surbanovski, PhD
	Mark Borodovsky, PhD
	Judson A Ward, PhD
	Kristoff Engelen, PhD
	Andrea Cavallini, PhD
	Alessandro Cestaro, PhD
	Daniel James Sargent, PhD
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	Dear Dr Zauner, please find attached our manuscript with the minor modifications made as you suggested. Many thanks for guiding our manuscript through the review process, it is greatly appreciated. Best regards, Dan Sargent
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<b>Experimental design and statistics</b>	Yes
<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
<b>Resources</b>	Yes
<p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	
<b>Availability of data and materials</b>	Yes

All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in [publicly available repositories](#) (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist?](#)

1 **The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to**  
2 ***Fragaria vesca* (the woodland strawberry)**

7 **DATA NOTE**

5 Matteo Buti<sup>1,§</sup> ([mbuti78@gmail.com](mailto:mbuti78@gmail.com)), Marco Moretto<sup>1</sup> ([marco.moretto@fmach.it](mailto:marco.moretto@fmach.it)), Elena Barghini<sup>2</sup>  
([elena.barghini@gmail.com](mailto:elena.barghini@gmail.com)), Flavia Mascagni<sup>2</sup> ([flaviamascagni@gmail.com](mailto:flaviamascagni@gmail.com)), Lucia Natali<sup>2</sup>  
([lucia.natali@unipi.it](mailto:lucia.natali@unipi.it)), Matteo Brilli<sup>1,3†</sup> ([matteo.brilli@unimi.it](mailto:matteo.brilli@unimi.it)), Alexandre Lomsadze<sup>4</sup>  
([alexandre.lomsadze@bme.gatech.edu](mailto:alexandre.lomsadze@bme.gatech.edu)), Paolo Sonogo<sup>1</sup> ([paolo.sonogo@fmach.it](mailto:paolo.sonogo@fmach.it)), Lara Giongo<sup>1</sup>  
([lara.giongo@fmach.it](mailto:lara.giongo@fmach.it)), Michael Alonge<sup>5</sup> ([malonge11@gmail.com](mailto:malonge11@gmail.com)), Riccardo Velasco<sup>1</sup>  
([riccardo.velasco@crea.gov.it](mailto:riccardo.velasco@crea.gov.it)), Claudio Varotto<sup>1</sup> ([claudio.varotto@fmach.it](mailto:claudio.varotto@fmach.it)), Nada Šurbanovski<sup>1</sup>  
([surbanovski.nada@gmail.com](mailto:surbanovski.nada@gmail.com)), Mark Borodovsky<sup>3</sup> ([borodovsky@gatech.edu](mailto:borodovsky@gatech.edu)), Judson A. Ward<sup>4</sup>  
([judson.ward@driscolls.com](mailto:judson.ward@driscolls.com)), Kristof Engelen<sup>1</sup> ([engelen.kristof@gmail.com](mailto:engelen.kristof@gmail.com)), Andrea Cavallini<sup>2</sup>  
([andrea.cavallini@unipi.it](mailto:andrea.cavallini@unipi.it)), Alessandro Cestaro<sup>1</sup> ([alessandro.cestaro@fmach.it](mailto:alessandro.cestaro@fmach.it)), Daniel James  
Sargent<sup>1,6,\*</sup> ([sargentdj@gmail.com](mailto:sargentdj@gmail.com))

<sup>1</sup>Fondazione Edmund Mach, Centre for Research and Innovation, via Mach 1, San Michele  
all'Adige, 38010 (TN), Italy

<sup>§</sup>Present address: Center for the Development and Improvement of Agri-Food Resources (BIOGEST-  
SITEIA) University of Modena and Reggio Emilia, P.le Europa 1, 42124 Reggio nell'Emilia (RE),  
Italy

<sup>2</sup>Department of Agricultural, Food, and Environmental Sciences, University of Pisa, Pisa I-56124,  
Italy.

<sup>3</sup>Department of Agronomy, Food, Natural Resources, Animals and Environment, University of  
Padova Agripolis, V.le dell'Università 16, 35020 Legnaro (PD), Italy.

<sup>†</sup>Present address: Dipartimento di Bioscienze e Centro di Ricerca Pediatrica Romeo ed Enrica  
Invernizzi, Università degli Studi di Milano, Via Celoria 26, 20133 Milano.

1 <sup>4</sup>Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA 30332,  
2 USA.

3  
4 <sup>5</sup>Driscoll's Strawberry Associates, Cassin Ranch, 121 Silliman Drive, Watsonville, California,  
5 USA.

6  
7 <sup>6</sup>Driscoll's Genetics Limited, East Malling Enterprise Centre, New Road, East Malling, Kent ME19  
8 6BJ, UK.

9 \*Corresponding Author

## 10 **ABSTRACT**

11 **Background:** The genus *Potentilla* is closely related to that of *Fragaria*, the economically important  
12 strawberry genus. *Potentilla micrantha* is a species that does not develop berries, but shares numerous  
13 morphological and ecological characteristics with *F. vesca*. These similarities make *P. micrantha* an  
14 attractive choice for comparative genomics studies with *F. vesca*. **Findings:** In this study, the  
15 *Potentilla micrantha* genome was sequenced and annotated, and RNA-Seq data from the different  
16 developmental stages of flowering and fruiting were used to develop a set of gene predictions. A 327  
17 Mbp sequence and annotation of the genome of *P. micrantha*, spanning 2,674 sequence contigs, with  
18 an N50 size of 335,712, estimated to cover 80% of the total genome size of the species was developed.  
19 The genus *Potentilla* has a characteristically larger genome size than *Fragaria*, but the recovered  
20 sequence scaffolds were remarkably collinear at the micro-syntenic level with the genome of  
21 *F. vesca*, its closest sequenced relative. A total of 33,602 genes were predicted, and 95.1% of BUSCO  
22 genes were complete within the presented sequence. Thus, we argue that the majority of the gene-  
23 rich regions of the genome have been sequenced. **Conclusions:** Comparisons of RNA-Seq data from  
24 the stages of floral and fruit development revealed genes differentially expressed between  
25 *P. micrantha* and *F. vesca*. The data presented are a valuable resource for future studies of berry  
26 development in *Fragaria* and the Rosaceae and they also shed light on the evolution of genome size  
27 and organization in this family.

1  
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

*Keywords:* long-read sequencing; evolutionary development; angiosperms; genome sequence; transcriptomics;

**BACKGROUND**

*Potentilla*, a genus of approximately 500 species [1], is closely-related to that of *Fragaria* [2], the genera having diverged from a common ancestor just 24 million years ago [3]. The genus *Fragaria*, a member of the Fragariinae tribe of the Rosaceae family, is economically-important due to the sweet, aromatic accessory fruits (berries) produced by members of the genus, in particular those of the cultivated allo-octoploid ( $2n=8\times=56$ ) strawberry species *F. × ananassa*. The availability of a genome sequence for a wild diploid relative of the cultivated strawberry, the woodland strawberry *F. vesca* ( $2n=2\times=14$ ) [4] has enabled the investigation of the molecular basis of many traits of economic and academic interest in strawberry, including the development of accessory fruits. However, all members of the *Fragaria* genus produce berries, and as such the use of reverse genetics approaches to study the genes involved in berry evolution and development would require *Fragaria* mutants that do not produce fruits, a resource that is not currently available.

In the post genomics era comparative analysis permits the study of related, yet divergent species, by tracing changes at the genomic and transcriptomic levels responsible for their phenotypic differences. Previously, the sequenced genomes of *F. vesca*, *Prunus persica* and *Malus × domestica* were compared [5], providing insights into the evolutionary mechanisms that have shaped the three species, and demonstrating that the *Fragaria* genome underwent significant small-scale structural rearrangements since it diverged from the common ancestor of the three genera. Comparative transcriptomics can also be used to reveal differences in the expression of orthologous genes between organisms at different stages of physiological development [6]. Such an approach suggests that comparative analyses between *Fragaria* and a closely-related species that does not bear berries may reveal important insights into the evolution of fruit development. Additionally, speciation is often

1 related to changes in genome structure, and genome size in particular. Differences in genome size are  
2 often the consequence of polyploidization events and/or changes in the abundance of repetitive DNA,  
3 especially transposable elements [7].

4  
5 *Potentilla micrantha*, like the majority of species of the genus *Potentilla* does not develop accessory  
6 fruits, but it shares numerous morphological characteristics with *F. vesca* (Fig. 1) including plant  
7 habit and flower morphology. Notably, they grow within the same ecological niches, and where their  
8 ranges of distribution overlap, *P. micrantha* can be found growing nearby populations of *F. vesca*  
9 (Sargent, unpublished results). These striking similarities make *P. micrantha* an attractive choice for  
10 understanding the genetic basis of berry development in *F. vesca*. As a precursor to a whole genome  
11 sequencing initiative, an initial sequencing project focused on the *P. micrantha* chloroplast was  
12 undertaken using the Illumina HiSeq and PacBio RS sequencing platforms [8].

## 13 **DATA DESCRIPTION**

14 The objectives of this study were to develop a genomic toolkit for *P. micrantha* to permit comparative  
15 genomic and transcriptomic studies with *F. vesca*, with a view to identifying the evolutionary changes  
16 that have occurred between the two species. The genome size of *P. micrantha* was determined by  
17 flow cytometry and the nuclear genome was sequenced and assembled from Illumina and PacBio  
18 sequencing reads, assembled and integrated using ALLPATHS and PBJelly. Gene predictions from  
19 the *P. micrantha* genome were made with support of RNA-Seq data generated from tissue libraries  
20 sampled during flower and fruit development. The genome of *F. vesca* was compared to the  
21 sequencing scaffolds produced for *P. micrantha*, and whilst they exhibited a remarkable degree of  
22 collinearity at the micro-syntenic level, large-scale differences in transposon activity were identified  
23 that might explain the large differences in genome size between the two species. The dataset we report  
24 will be useful for comparative studies of a number of traits between *P. micrantha* and its  
25 economically-important close relatives.

## 1 Flow cytometry, heterozygosity estimation and genome assembly

2 DNA was extracted from *Potentilla micrantha* young, unexpanded leaves. Flow cytometry using a  
3  
4  
5 *V. minor* internal standard with a DNA content of 1.52 pg/2C returned average DNA quantities of  
6  
7 0.52 pg/2C for *F. vesca* ‘Hawaii 4’ and 0.83 pg/2C for *P. micrantha* over three biological replicates.  
8  
9  
10 Using the calculation of [9] that 1 pg DNA is equivalent to 978 Mbp of DNA sequence, the genome  
11  
12 size of *P. micrantha* was determined as 405.87 Mbp in length whilst that of *F. vesca* ‘Hawaii 4’ was  
13  
14 calculated to be 254.28 Mbp.  
15

16  
17 Data were returned for the overlapping fragment library (OLF) and all four mate-pair libraries  
18  
19 sequenced using Illumina HiSeq. In total, 61.4 Gbp of data were returned and the relative depth of  
20  
21 coverage obtained for the *P. micrantha* genome from each library is given in Additional File 1: Table  
22  
23  
24 S1. Four different PacBio RS sequencing libraries were constructed and sequenced using two  
25  
26 different versions of the PacBio chemistry (Additional File 2: Table S2). From the sequencing of 63  
27  
28 SMRT cells, 6,447,413 sequences with an average length of 2,221 bp were recovered, totaling  
29  
30  
31 14.32 Gb of long read sequence data. From the data, 33× equivalent of sequence was contained in  
32  
33  
34 reads longer than 1 kb which were used for gap filling of the Illumina assembly using PBJelly [10].  
35

36 The initial ALLPATHS assembly of the Illumina short-read sequences produced 33,026 contigs with  
37  
38 an N50 of 16,235 bp and a total length of 247,565,733 bp. Following scaffolding, a genome assembly  
39  
40  
41 with a total length of 315,266,043 bp contained in 2,866 sequencing scaffolds was returned. The final  
42  
43 scaffold set returned following ALLPATHS assembly contained a total of 0.07% ambiguous sites  
44  
45 (SNPs), revealing the genome of *P. micrantha* to be one of the most homozygous naturally-occurring  
46  
47 genomes sequenced to date. Following incorporation of the PacBio RS data using PBJelly [10], the  
48  
49  
50  
51 *P. micrantha* sequence assembly contained 326,533,584 bp of sequence data, a 3.5% increase over  
52  
53 the ALLPATHS Illumina assembly, in 2,674 scaffolds. The longest and N50 scaffold lengths both  
54  
55 increased following gap filling by 9.3% and 5.1% respectively, but most significantly, the number of  
56  
57 gapped Ns in the assembly was reduced by 59.7% to 27,311,787 (8.4% of the final assembly) (Table  
58  
59  
60  
61 1). The final scaffolded assembly contained 80.45% of the total estimated genome size for  
62  
63  
64  
65



1 *P. micrantha* as calculated by flow cytometry. Scaffolds ranged from 935 bp to 3,488,351 bp in  
2 length. Of the 2,674 scaffolds, 878 (32.8%) were less than 10 kbp in length, 534 (20%) were between  
3  
4 10 and 50 kbp in length, 738 (27.6%) were between 50 and 200 kbp in length, 500 (18.7%) contained  
5  
6 between 200 kbp and 1 Mbp of sequence, and the remaining 23 (0.9%) contained over 1 Mbp of  
7  
8 sequence. The majority of the 1,440 benchmarking single-copy orthologous (BUSCO) groups queried  
9  
10 [11] were present in the genome sequence, with 95.1% (1,337 complete and single copy and 33  
11  
12 complete and duplicated BUSCOs) identified within the sequencing scaffolds.  
13  
14  
15  
16  
17  
18

### 19 **Gene prediction and preliminary annotation**

20  
21 The results of the combined alignment of the 12 RNA-seq read sets to the *P. micrantha* genome  
22  
23 assembly and number of splice sites identified using STAR is presented in Additional File 3: Table  
24  
25 S3. A total of 1,908 consensus repeat sequences were generated by RepeatModeler totaling  
26  
27 1,431,262 bp and having a GC content of 40.8%. The total ATCG content of sequencing scaffolds  
28  
29 greater than 10 kb in length was 298,987,576 bp. A total of 138,597,969 bp (46.36%) of the genome  
30  
31 sequence were masked using the consensus sequences in the RepeatModeler library, including 26,359  
32  
33 (7.5%) of the mapped GT-AG introns identified by STAR. Gene prediction using GeneMark-ET on  
34  
35 the masked genome identified a total of 33,602 genes, of which 32,137 were predictions containing  
36  
37 multiple exons, and 4,655 were single exon predictions. A total of 172,791 exons were predicted,  
38  
39 with an average length of 223 bp and an average of 5.14 exons per gene. A total of 139,216 introns  
40  
41 were predicted in the CDS of the genes, with an average intron length of 499 bp. BUSCO analyses  
42  
43 were compared between the gene predictions developed for *P. micrantha* and those of *F. vesca*. In  
44  
45 total, 1,282 (89%) complete and 68 (4.7%) fragmented BUSCOs (93.75% total) were recovered for  
46  
47 *P. micrantha*, compared to 1,303 (90.5%) complete and 79 (5.5%) fragmented BUSCOs (95.6%)  
48  
49 recovered for *F. vesca* gene predictions indicating a similar level of completeness of the *P. micrantha*  
50  
51 assembly to its nearest sequenced relative. Following a local BLAST search and BLAST2GO  
52  
53 analysis, a total of 27,968 *P. micrantha* predicted genes were assigned a preliminary gene annotation.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## Scaffold anchoring and synteny to the *Fragaria vesca* Fvb genome sequence

Following the inparanoid analysis, a total of 33,127 genes returned an orthologous relationship with one or more *F. vesca* gene predictions at the amino acid level (98.6%). A subsequent BLAST analysis of the gene predictions against the *F. vesca* v2.0 pseudomolecules identified a total of 24,641 *P. micrantha* genes that returned an unambiguous match with a *F. vesca* orthologue. A total of 1,682 *P. micrantha* sequence scaffolds, containing 315,081,089 bp (96.5% of the total sequence) contained at least one gene that was anchored to one of the *F. vesca* v2.0 pseudomolecules. Of those, 573 contained at least ten orthologous gene sequences, 118 contained at least 50 orthologous sequences and 32 contained over 100 orthologous (Supplementary Excel File 1). Scaffold ‘Contig145’, the largest scaffold in the *P. micrantha* genome sequence (3,488,351 bp) contained the largest number of orthologous gene sequences anchored to the *F. vesca* v2.0 genome sequence (560), whilst scaffold ‘Contig2191’ was the smallest anchored scaffold at 1,163 bp, and containing a single orthologous gene sequence. Comparison of the two genomes revealed a remarkable degree of micro-synteny with the majority of the *P. micrantha* scaffolds spanning uninterrupted regions of the *F. vesca* genome sequence (Data not shown). A very high degree of collinearity in gene order was observed between *P. micrantha* scaffolds and the *F. vesca* pseudomolecules (Fig. 2a). In general, only a small number of inversions were observed between syntenic blocks between the two genomes, and just eight *P. micrantha* scaffolds contained distinct syntenic blocks that aligned with more than one *Fragaria* pseudomolecule (Fig. 2b). However, scaffold anchoring to a genetic map however was not performed for the *P. micrantha* genome sequence, and as such, a comparison of macrosynteny between *Fragaria* and *Potentilla* could not be made.

## Gene expression during fruit development

Tissues from five stages of flowering and ‘fruit’ development were harvested from *P. micrantha* flowers in biological duplicates or triplicates for RNA isolation. The stages of flowering followed

1 those identified in *Fragaria* by [12], with the addition of a stage 0 (unopened flowers) and young  
2 unexpanded leaf tissue. The selected developmental stages are shown in Fig. 3. RNA-libraries were  
3 made and sequenced with Illumina HiSeq2000. Following QC and adapters trimming, a total of  
4 619,085,115 101 bp paired reads were obtained from the 12 *P. micrantha* RNA-seq libraries.  
5 Sequencing yield from individual libraries ranged from 29,653,058 to 60,158,302 reads per sample  
6 (Additional File 4: Table S4). Following trimming, the number of reads available for *Fragaria* from  
7 the published sequences of [12] were 1,236,882,540, with reads per library ranging from 109,643,225  
8 to 155,643,061. Between 62% and 69% of *P. micrantha* filtered reads per library mapped to the  
9 *P. micrantha* gene prediction set, and 63% to 67% of *F. vesca* filtered reads per library mapped to  
10 the *F. vesca* gene predictions (Additional File 4: Table S4). A total of 1,556 genes were differentially  
11 expressed between the four developmental stages in at least one pair-wise comparison of the different  
12 stages in *P. micrantha*, whilst in *F. vesca*, 816 genes were differentially expressed in at least one of  
13 the contrasts (Fig. 4). A total of 52.44% and 43.38% differentially expressed genes were GO-  
14 annotated for *P. micrantha* and *F. vesca* respectively (Additional File 5: Fig. S1). Analysis of the GO  
15 terms for *F. vesca* and *P. micrantha* revealed an enrichment for lipid metabolic processes, transporter  
16 activity, and transcription factor activity and transcription regulator activity in *F. vesca* over  
17 *P. micrantha* (Fig. 5). The gene expression profiles between the four developmental stages studied in  
18 the two species showed no clear consistent patterns between the two species overall (Additional File  
19 6: Fig. S2), however the common differentially expressed genes displayed largely similar expression  
20 patterns (Fig. 6), with some exceptions, most notably gene1369-v1.0-hybrid and its homologue in  
21 *P. micrantha* (17717\_t), a predicted 3-hydroxy-3-methylglutaryl coenzyme A reductase 1, which was  
22 highly expressed in *F. vesca* but exhibited far lower levels in *P. micrantha*.

### 23 **Analysis of MADs-box conserved domain-containing genes in *Potentilla* and *Fragaria***

24 A total of 75 *P. micrantha* and 81 *F. vesca* predicted proteins containing MADS-box conserved  
25 domains were aligned and phylogenetic trees were obtained to reliably identify orthology

1 relationships between *P. micrantha* and *F. vesca* genes. The three methods employed for phylogenetic  
2 reconstruction (ML, MP, NJ) returned largely congruent topologies for the nodes with more than 50%  
3 bootstrap support, with NJ providing a slightly more resolved tree given the use of a pairwise, instead  
4 of a partial deletion approach. Fig. 7 displays the ML phylogenetic reconstruction of the *P. micrantha*  
5 and *F. vesca* genes containing MADs-box, along with the gene expression levels for each gene (data  
6 for the NJ and MP trees are not shown). The majority of the genes were retained after the divergence  
7 of the species, indicated by a large proportion of orthologous pairs retrieved. Only a few events of  
8 lineage-specific gene loss/duplication were observed. Both observations are in line with the lack of  
9 ploidy changes within *P. micrantha* and *F. vesca* in the estimated 24.22 million years since species  
10 divergence. As expected, the majority of orthologous pairs shared similar expression patterns. Based  
11 on the ML gene tree however, three clades of orthologous genes were identified that were not  
12 expressed, or poorly expressed in *P. micrantha* but highly expressed in *F. vesca* (Fig. 8). The three  
13 clades, numbered as 1, 2 and 3 on Fig. 8, contained the following genes: clade 1 contained genes  
14 27280\_t (*P. micrantha*) and gene25871-v1.0-hybrid (*F. vesca*), which displayed highest homology to  
15 *A. thaliana* AGL36, a sequence-specific DNA binding transcription factor active during endosperm  
16 development [13]; clade 2 contained genes 26598\_t (*P. micrantha*) and gene18483-v1.0-hybrid  
17 (*F. vesca*), whose closest *A. thaliana* homologue was AGL62, a MADS gene that promotes embryo  
18 development, indicating an essential role of endosperm cellularization for viable seed formation [14];  
19 and clade 3 contained *P. micrantha* genes 23638\_t, 23641t and 759\_t and *F. vesca* genes gene32155-  
20 v1.0-hybrid and gene13277-v1.0-hybrid, whose closest *A. thaliana* homologue AGL15 delays  
21 senescence programs in perianth organs and developing fruits and alters the process of seed  
22 desiccation [15].

### 23 **Analysis of the repetitive component of the *Potentilla micrantha* genome**

24 In total, 1,001,838 of 1,484,780 reads clustered with RepeatExplorer were grouped into 107,190  
25 clusters, representing 67.5% of the genome. No predominant repeat families were identified in the  
26

1 *P. micrantha* genome, with the most redundant repeat cluster representing just 1.18% of the total  
2 genome length. LTR-retrotransposons made up the main fraction (24.1%) of the *P. micrantha*  
3 genome (Additional File 7: Fig. S3), with a *Gypsy* to *Copia* ratio of approximately 2:1. Terminal-  
4 repeat retrotransposons in miniature (TRIMs) were poorly represented, making up just 0.2% of the  
5 genome, whilst putative DNA transposons accounted for 5.7% of the genome and included putative  
6 CACTA, Harbinger, and hAT elements, with other, unclassified repeats accounting for 10.6% of the  
7 genome. A comparison of the repetitive portion of the *F. vesca* and *P. micrantha* genomes performed  
8 by pairwise clustering of Illumina sequence reads revealed significant diversification between the  
9 repetitive component of the genomes of the two species (Additional File 8: Fig. S4). Among the top  
10 291 repeat clusters that had a genome proportion >0.01%, 107 were specific to *P. micrantha*, 51 were  
11 specific to *F. vesca*, whilst only 25 were similarly represented in the two species. Among all repeat  
12 classes, only ribosomal DNAs show similar genome proportions between *P. micrantha* and *F. vesca*.

#### 31 ***Potentilla* full-length LTR-RE characterization, annotation and insertion age**

32 Of the 505 characterised LTR-REs, 220 (43.6%) belonged to the *Copia* superfamily, with the greatest  
33 proportion belonging to the *Bianca* family, 256 (50.7%) belonged to the *Gypsy* superfamily, with the  
34 greatest proportion belonging to the *Ogre/TAT* family, whilst the remaining 29 (5.7%) could not be  
35 placed into a specific superfamily. Table 2 lists the proportion of the annotated 505 LTR-REs in each  
36 superfamily, and the numbers of elements contained in each sub-family within the *Copia* and *Gypsy*  
37 super-families. For RE insertion age determination, a mean synonymous substitution rate between  
38 *P. micrantha* and *F. vesca* of 0.064 ( $K_s$ ), was estimated by comparing 50 orthologous genes, which  
39 equated to 52,703 bp of aligned sequences. Using a timescale of 24.22 million years since the  
40 separation of *P. micrantha* and *F. vesca*, and the estimated  $K_s$  of 0.064, a synonymous substitution  
41 rate of  $2.64 \times 10^{-9}$  substitutions per year was calculated. As mutation rates for LTR retrotransposons  
42 have been estimated to be approximately two-fold higher than silent site mutation rates for protein  
43 coding genes [16,17], a substitution rate per year of  $5.28 \times 10^{-9}$  was used in calculations of LTR-RE

1 insertion dates. When the whole set of usable retrotransposons was taken into account, the nucleotide  
2 distance (K) between sister LTRs showed a large degree of variation between retro-elements, ranging  
3  
4 from 0 to 0.124 using the Kimura two parameter method, which represents a time span of at most  
5  
6  
7 23.54 million years.  
8  
9

## 10 **DISCUSSION**

### 11 **Data validation and quality control**

12 In this investigation, the genome of *P. micrantha*, a member of the Rosaceae, a diverse family of  
13  
14 fruiting perennial plant genera, was sequenced using both short-read Illumina and long-read PacBio  
15  
16 sequence data, and the resulting data was assembled into a highly contiguous reference sequence for  
17  
18 the genus *Potentilla*. PacBio data (using early iterations of the sequencing chemistry) were  
19  
20 proficiently integrated with short-reads, significantly improving the contiguity of the assembly. The  
21  
22 genome assembly presented here has a quality similar to the *F. vesca* genome, containing  
23  
24 significantly fewer un-sequenced gaps within scaffolds, and is far more contiguous than that of *R.*  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000

1 *F. vesca* has been previously demonstrated to contain just 22% repetitive elements [4]. Contrary to  
2 the coding or non-repetitive genome, the repetitive fractions of the *P. micrantha* and *F. vesca*  
3 genomes are highly diversified, suggesting that the overwhelming majority of retrotransposon activity  
4 in the genus *Potentilla* occurred after the divergence of the two genera from their common progenitor.  
5 The data presented here strongly indicate that retrotransposon activity (or the lack thereof in the genus  
6 *Fragaria*) is responsible for the significant difference between the genome size of *Fragaria* and its  
7 closest relatives, and support the assertion of [2] that *Fragaria* should be treated as a distinct genus,  
8 separate from *Potentilla*.  
9 Gene expression patterns for differentially expressed genes that were common to both *F. vesca* and  
10 *P. micrantha* were largely similar between the two species, however one gene, a 3-hydroxy-3-  
11 methylglutaryl coenzyme A reductase 1 homologue displayed significantly higher gene expression  
12 levels in *F. vesca*. The 3-hydroxy-3-methylglutaryl coenzyme A reductase 1 gene catalyzes the first  
13 committed step in the cytosolic isoprenoid biosynthesis pathway [24]. Loss of function mutants of  
14 this gene in *Arabidopsis* display a dwarf phenotype due to suppression of cell elongation and reduced  
15 sterol levels [24]. Sterols are precursors in cellulose synthesis, important for cell-wall formation [25]  
16 and fruit development, and as such, up-regulation in the 3-hydroxy-3-methylglutaryl coenzyme A  
17 reductase 1 gene during fruit development in *F. vesca* over *P. micrantha* may indicate a role for this  
18 enzyme in berry formation in *Fragaria*.  
19 In contrast to the gene expression patterns of differentially expressed genes common to both *F. vesca*  
20 and *P. micrantha* during fruit development, global patterns of gene expression during fruit  
21 development differed between the two species. The gene ontology for the *F. vesca* expression profile  
22 was enriched for genes with transcription factor and transcription regulator activity as well as  
23 transporter activity and lipid metabolic processes. A study of the differences in transcriptional  
24 regulation between *F. vesca* and *P. micrantha* therefore may provide clues to the genetic basis of  
25 berry formation in *F. vesca*. MADS-box transcription factors have been implicated in a wide and  
26 extremely diverse array of developmental processes in plants [26], and were initially demonstrated to

1 play a major role in floral organ differentiation, including gametophyte, embryo and seed  
2 development, as well as flower and fruit development. A study of the differential expression of  
3  
4 MADS-box genes revealed three clades of orthologous genes where gene expression of orthologous  
5  
6 genes was up-regulated in *F. vesca* with respect to *P. micrantha*, where the genes were either shown  
7  
8 to have lower expression levels, or were not expressed in the tissues studied. One clade contained  
9  
10 genes that were homologous to AGL36, a transcription factor crucial for endosperm differentiation  
11  
12 and development [13,27]. Another clade contained genes homologous to *A. thaliana* AGL62, which  
13  
14 likewise has been implicated in embryo development, and is thought to have an essential role of  
15  
16 endosperm cellularization for viable seed formation [14]. The third clade contained genes  
17  
18 homologous to AGL15 reported to have diverse roles in embryogenesis, fruit maturation, seed  
19  
20 desiccation and the repression of floral transition [15,28], as well as being a positive regulator of the  
21  
22 expression of mir156, a repressor of floral transition [29].  
23  
24  
25  
26  
27  
28

### 29 **Re-use potential**

30  
31 The set of genomics tools developed here for *P. micrantha*, a non-fruitlet relative of *F. vesca* includes  
32  
33 a genome sequence, gene predictions and RNA-Seq data. It is a valuable resource and will form the  
34  
35 foundation for future genomics studies in the species and comparative genomics studies within the  
36  
37 Rosoideae sub-family of Rosaceae in particular. It will also allow more detailed future functional  
38  
39 studies of fleshy receptacle (berry) development.  
40  
41  
42  
43  
44  
45

## 46 **METHODS**

### 47 **Plant material, flow cytometry and DNA isolation**

48  
49 A specimen of *P. micrantha* was collected from Avala, Serbia in spring 2012 and subsequently used  
50  
51 for sequencing. The plant was maintained in a growth room at a constant temperature of 24 degrees  
52  
53 during the day and 18 degrees at night, with a 16-hour photoperiod to encourage new shoot  
54  
55 development. Young leaves were harvested and subjected to flow cytometry by Plant Cytometry  
56  
57 Services, NL. Measurements were taken in triplicate against a *Vicia minor* internal standard using the  
58  
59  
60  
61  
62  
63  
64  
65



1 propidium iodide fluorescent dye. The *F. vesca* accession ‘Hawaii 4’ for which a whole genome  
2 sequence has been published [23] was analyzed for comparison. Prior to harvesting leaf material for  
3 DNA extraction, the plant was moved to a darkened growth chamber for 120 hours, maintaining a  
4 constant temperature of 22 degrees. DNA was extracted from young, unexpanded leaf material using  
5 the modified CTAB extraction protocol [30], quantified using a Nanodrop spectrophotometer and  
6 Qubit fluorometer, and assessed for integrity by agarose gel electrophoresis against a  $\lambda$  *Hind*III size  
7 standard.

8 Since *P. micrantha* does not reproduce asexually from runners, a seedling population obtained from  
9 the selfing of the original mother plant was maintained from which to harvest tissue from stages of  
10 floral and fruiting development. Flowers of *P. micrantha* and *F. vesca*, along with two other  
11 *Potentilla* species, *P. reptans* and *P. indica* were treated with naphthaleneacetic acid (NAA; Sigma-  
12 Aldrich), N-1-naphthylphthalamic acid (NPA; Sigma-Aldrich), gibberellic acid (GA3; Sigma-  
13 Aldrich) and a combination of NAA and NPA, following the methods of [12]. Briefly, stock solutions  
14 of 50 mM NAA, 50mM NPA, and 100mM GA3 were made in ethanol and diluted with two drops of  
15 Tween 20 and water before application. The final treatment concentrations were 500  $\mu$ M for NAA  
16 and GA3 and 100  $\mu$ M for NPA. 50 ml of hormone solution was pipetted onto the receptacle of each  
17 emasculated flower every two days for twelve days.

### 18 **Tissue sampling, RNA extraction and sequencing**

19 Tissues from five stages of flowering and ‘fruit’ development were harvested from untreated flowers  
20 in biological duplicates or triplicates for RNA isolation. The stages of flowering followed those  
21 identified in *Fragaria* by [12], with the addition of a stage 0 (unopened flowers) and young  
22 unexpanded leaf tissue. The selected developmental stages are shown in Fig. 3. RNA was extracted  
23 from 50 mg of snap-frozen tissue from each developmental stage using the Spectrum plant total RNA  
24 extraction kit (Sigma) with an on-column DNase I digestion (Sigma) step. The extraction protocol  
25 followed the manufacturers’ recommendations with two minor modifications: 1% PVP was added to

1 the lysis solution, and the number of washes at each stage was doubled (i.e. two washes were  
2 performed with wash solution 1 and four washes were performed with wash solution 2). The RNA  
3 extracted from each sample was diluted in 50 µl of elution solution (Sigma). Following elution, total  
4 RNA was quantified using a Nanodrop spectrophotometer and Qubit fluorometer and assessed for  
5 integrity using a Bioanalyzer (Agilent). Samples returning a RIN value greater than 7.5 were  
6 considered acceptable for sequencing. A total of 12 Illumina TruSeq libraries were constructed from  
7 2 µg of total RNA. Libraries were made from the following samples; one from stage 0, two from  
8 stage 1, two from stage 2, three from stage 3 and three from stage 4. A final library was made from  
9 RNA of young leaf tissue. The libraries were sequenced in triplex per single lane of Illumina  
10 HiSeq2000. Samples were indexed and multiplexed, and then 101 bp paired-end sequencing was  
11 performed using the Illumina HiSeq 2000 platform at the Weill Medical core genomics facility of  
12 Cornell University.

### 13 **Whole genome shotgun sequencing, assembly**

14 A strategy following the ALLPATHS-LG protocol was followed to produce an initial assembly using  
15 second-generation sequence data. Five sequencing libraries were developed; an overlapping fragment  
16 library (OLF) with an insert size of 170 bp, and four libraries of 3 kb, 5 kb, 8 kb and 12 kb. The OLF  
17 library was created using the Illumina Nextera library preparation kit following the manufacturers'  
18 recommendations and was sequenced in simplex on a single lane of Illumina HiSeq2000, whilst the  
19 MP libraries were prepared using the Illumina Mate Pair Library v2 kit following the manufacturers'  
20 recommendations and were subsequently sequenced in duplex. All sequencing was performed at the  
21 Weill Medical Centre core genomics facility at Cornell University. ALLPATHS-LG (ALLPATHS-  
22 LG, RRID:SCR\_010742) [31] was run using the sequencing libraries described above using default  
23 settings. Subsequently, a selection of SMRT-bell sequencing libraries were constructed using various  
24 versions of the PacBio RS sequencing kits and chemistries (Additional File 2: Table S2) and PBJelly  
25 (PBJelly, RRID:SCR\_012091) [10] running default settings was used to incorporate data generated

1 using the PacBio RS platform (Pacific Biosciences) into the ALLPATHS-LG Illumina assembly  
2 scaffolds. Identification of benchmarking universal single-copy orthologs was performed using  
3  
4 BUSCO v3 (BUSCO, RRID:SCR\_015008) [11] running default parameters and using 1,440 BUSCO  
5  
6  
7 groups from the embryophyta\_odb9 (plant) lineage data.  
8  
9

## 10 **Gene prediction, annotation, determination of gene orthology and evaluation of synteny** 11 12 **between *Potentilla* and *Fragaria* genomes** 13 14

15  
16 First, *ab initio* repeat finding was done with RepeatModeler (RepeatModeler, RRID:SCR\_015027)  
17  
18 [32] that was run on the complete set of genomic scaffolds set and a repeat library was created. Next,  
19  
20 the genome was masked using RepeatMasker (RepeatMasker, RRID:SCR\_012954) [33]. Gene  
21  
22 prediction was done with GeneMark-ET [34]. The following parameters were used; a minimum  
23  
24 scaffold length of 10 kb, a maximum scaffold gap size of 40 kb, a minimum intron size of 50 bp, a  
25  
26 maximum intron length of 10 kb and a maximum intergenic length of 50 kb. RNA-seq reads from the  
27  
28 12 libraries were aligned to the genome sequence scaffolds using the STAR tool with default  
29  
30 parameters [35]. Reads from the 12 RNA-seq datasets were aligned to the genome. Mapping of RNA-  
31  
32 seq reads that included intron junctions led to the identification of introns. Introns with a high ‘intron  
33  
34 score’ (identified by more than 60 RNAseq reads) were considered to be reliably identified. Predicted  
35  
36 genes were annotated using BLAST2GO (BLAST2GO, RRID:SCR\_005828) [36]. The non-  
37  
38 redundant NCBI protein database was downloaded and BLAST was run locally. Results from the  
39  
40 BLAST analysis were uploaded to the BLAST2GO server and gene ontology analyses were  
41  
42 performed using default parameters.  
43  
44

45  
46 Orthologous relationships between *Fragaria* and *Potentilla* genes was determined through sequence  
47  
48 clustering performed using Inparanoid 7 [37]. Analyses were based only on homology, as an  
49  
50 alternative to the more stringent ortholog classification. *Prunus persica* v2.0.a1 predicted proteins  
51  
52 downloaded from the GDR [38] and *P. micrantha* and *F. vesca* protein sequences were blasted all  
53  
54 against all and the output file was filtered at the following thresholds: maximum E-value= $10^{-4}$  and  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 query coverage of at least 50%. The resulting file was used as an input to the MCL algorithm using  
2 as edge weight  $-\log_{10}(\text{evaluate})$  (all E-values=0 were changed to 1E-300). To explore more thoroughly  
3 the homology network used as input, the MCL algorithm was run at different granularity levels  
4 (inflation parameter equal to 1.5, 1.7, 2.0, 2.3, 2.4, 2.7, 3) and then a table indicating cluster  
5 memberships at the different stringencies was compiled for each node. Ortholog classification was  
6 produced using Inparanoid 7 [37] for pairs of species in all combinations. The resulting sqltables were  
7 then used as an input for QuickParanoid [39] and the sequences were combined in a three-species  
8 ortholog classification. The clusters obtained with QuickParanoid were used to calculate the number  
9 of genes contained in each cluster for both *Potentilla* and *Fragaria*.

10 *Potentilla* gene predictions for which an orthologous relationship was identified through the  
11 inparanoid analysis, were used as queries to identify the physical locations of orthologous sequences  
12 on the *F. vesca* v2.0 pseudomolecules and those sequences that returned a single, unambiguous match  
13 on the genome sequence were used to evaluate synteny between the two species. Since the *Potentilla*  
14 genomic scaffolds were not oriented and ordered against a reference genetic map, conservation of  
15 synteny between the *Potentilla* and *Fragaria* genomes was determined through a comparison of the  
16 physical positions of orthologous gene sequences on the sequence scaffolds of *Potentilla* and the  
17 pseudomolecules of *Fragaria*. Criteria for the identification of syntenic regions followed that of [5].  
18 No attempt was therefore made to infer macro-syntenic structure on a chromosome scale between the  
19 two genomes.

## 20 **Gene expression during stages of fruit development in *Potentilla micrantha* and *Fragaria vesca***

21 The quality of the raw reads generated as described above was checked with FastQC (FastQC,  
22 RRID:SCR\_014583) [40]; Trimmomatic (Trimmomatic, RRID:SCR\_011848) [41] was used to  
23 remove adapter sequences. The *F. vesca* .sra files [12] were used to compare gene expression in  
24 *Fragaria* with *Potentilla*; *Fragaria* reads from the same developmental stage were merged and treated  
25 as a single data set since data from *Potentilla* was not generated from individual floral organs. The

1 12 trimmed *P. micrantha* RNA-seq libraries were mapped on the *P. micrantha* gene prediction CDS,  
2 while the ten *F. vesca* sets were mapped to the *F. vesca* v1.0 gene prediction CDS [4] downloaded  
3 from the GDR [38] using Bowtie2 [42] and default settings. The number of reads mapping to each  
4 gene for each RNA set was calculated from the .sam alignment files derived from Bowtie2.  
5 Counts of RNA-seq reads over transcripts were used to calculate the gene expression level in  
6 FPKM= $10^9 \cdot ER / (EL \times MR)$ , where ER was the number of mapped reads in the exons of a particular  
7 gene, EL was the sum of exon length in base pairs, and MR was the total number of mapped reads  
8 [43]. FPKM was used to distinguish expressed genes from inactive genes (those not returning any  
9 expression data) during the flower development in each species. Further, FPKM was used to define  
10 a set of highly expressed genes: Genes were considered as ‘highly-expressed’ if FPKM>1000. Genes  
11 that returned an FPKM<1000 in all samples were removed from further differential expression  
12 analysis. The retained differentially expressed genes were processed by performing a linear rescaling  
13 of the log<sub>2</sub>-counts, aligning the distributions for every sample at their distribution modes, followed  
14 by variance stabilization to ensure homoscedasticity. A one-way ANOVA was performed gene-by-  
15 gene on the rescaled log<sub>2</sub>-counts to detect changes in expression among different developmental  
16 phases. Differentially expressed genes (DEGs) were selected by setting cutoffs both on the p-values  
17 from the ANOVA F-tests, as well as on the magnitude of observed changes represented by the square  
18 root of the ANOVA MSR values (equivalent to using volcano plots for two-condition studies). Genes  
19 were considered differentially expressed if the  $\sqrt{\text{MSR}} > 2.00$  and p-value  $< 10^{-3}$ .  
20 Gene Ontology enrichment analysis of DEG sets of *Potentilla micrantha* and *Fragaria vesca* was  
21 carried out using Blast2GO 2.8.0 [44] with “Fisher’s exact test” method, considering as “enriched”  
22 the GO categories with FDR<0.05. *Potentilla micrantha* whole transcriptome functional annotation  
23 obtained in this work was used as background for *Potentilla* GO enrichment analysis, while the  
24 “InterPro GO for GeneMark hybrid transcripts” database downloaded from GDR website was used  
25 as background for *Fragaria vesca*. Cytoscape 3.5.1 (Cytoscape, RRID:SCR\_003032) [45] with the  
26 BiNGO 3.0.3 plugin was used for the GO-slim network visualization of enriched GO categories over

1 *Fragaria vesca* and *Potentilla micrantha* DEGs. For determination of over-representation, the  
2 Benjamini and Hochberg FDR-adjusted significance level cutoff was 0.05.  
3  
4

#### 5 **Phylogenetic and functional analysis of MADS-box domain-containing genes and gene** 6 **expression profile mapping**

7 Protein sequences of *Potentilla* (this publication) and *Fragaria* (Fvesca\_v1.0\_hybrid; [38]) were  
8 analysed on the NCBI conserved domain database [46]. All proteins containing a MADS-box domain  
9 were retrieved and the MADS-box extracted with Bedtools getfasta [47] using default parameters.

10 An initial sequence alignment was carried out using ClustalW and pairwise distances were calculated  
11 to eliminate outliers. A total of 16 sequences were removed from further analysis since they were too  
12 short and possessed incomplete N-terminal ends, indicating they were likely pseudogenes. The  
13 alignment used for phylogenetic analysis was constructed with SATé-II [48] and contained 156  
14 protein sequences (75 from *Potentilla* and 81 from *Fragaria*).

15 Three methods, Maximum Likelihood (ML), Maximum Parsimony (MP) and Neighbour-joining  
16 (NJ), each with 1,000 bootstrap replicates were employed for phylogenetic reconstruction of the  
17 MADS-box domain containing genes using Mega 7.0.14 [49]. Where missing data was present in the  
18 alignment, deletion of columns containing a fraction of missing data above 10% and 30% was  
19 performed for ML and MP methods. Pairwise deletion was instead used in the case of NJ, to maximise  
20 the phylogenetic information retained in the alignment. The ML topology was used as reference for  
21 further analysis.

22 The expression profiles of the genes containing a MADS-box were used to decorate the phylogenetic  
23 tree using iTOL v2 [50], allowing the identification of orthologous MADS-box gene pairs displaying  
24 differential gene expression profiles between *Potentilla* and *Fragaria*. Curated annotation of  
25 differentially expressed putative gene function was carried out using BLASTp homology searches of  
26 the TAIR database [51].

## Analysis of the repetitive component of *Potentilla* genome

To identify and characterize genomic repeats in the *P. micrantha* genome, a reduced set of 2,000,000 randomly selected genomic Illumina reads, corresponding to 0.57× of the *P. micrantha* genome were subjected to clustering using RepeatExplorer [52]. Among the clusters produced, the top clusters, with a genome proportion higher than 0.01%, were annotated using 0.2 as cutoff for cluster connection through mates. Clusters that were annotated as similar to phi-X174 were removed as contaminants. The output of RepeatExplorer was also used to prepare an in-house library containing all contigs belonging to clusters annotated by RepeatExplorer as long terminal repeat retrotransposons (LTR-REs) by similarity search against RepBase [53]. Subsequently, pairwise hybrid clustering between a random set of 1,431,114 Illumina reads derived from *P. micrantha* genomic DNA and 1,090,102 *F. vesca* genomic reads, each corresponding to 0.41× of the respective genomes was performed using RepeatExplorer [52].

## *Potentilla* full-length LTR-RE characterization

LTR-FINDER [54] was used to isolate putative full-length LTR-REs from 280 randomly-selected *Potentilla* genome sequence scaffolds and alignment boundaries were obtained by adjusting the ends of LTR-pair candidates using the Smith–Waterman algorithm. These boundaries were re-adjusted based on the occurrence of the following typical LTR-RE features: (a) the putative LTR-RE were flanked by the dinucleotides TG and CA at 5' and 3' ends respectively; (b) a target-site duplication (TSD) of 4–6 nt in length was present in the sequence; (c) a putative 15–18 nt primer binding site (PBS) complementary to a tRNA at the end of the putative 5'-LTR was present in the sequence; and (d) a 20–25-nt polypurine tract (PPT) just upstream of the 5' end of the 3' LTR was present in the sequence. Putative LTR-REs were manually validated using DOTTER [55], verifying the occurrence of LTRs, dinucleotides TG and CA at the 5' and 3' ends respectively, and TSDs. The validated LTR-REs were annotated using BLASTX and BLASTN querying the NCBI nr nucleotide and protein

1 NCBI databases and RepBase [53]. To limit false-positive detection, a fixed E-value threshold of E  
2 <  $10^{-5}$  for BLASTN and  $E < 10^{-10}$  for BLASTX was used. The full-length elements identified were  
3  
4 analysed using RepeatExplorer [52], performing searches for GAG, protease, retrotranscriptase,  
5  
6 RNaseH, integrase, and chromodomain derived from plant protein domains from RepBase. The  
7  
8 similarity search was filtered at E-value  $< 10^{-10}$ , allowing for both mismatches and frameshifts. The  
9  
10 same tool was used to assign full-length elements to specific *Gypsy* or *Copia* lineages. Full-length  
11  
12 LTR-REs that were identified as belonging to *Gypsy* or *Copia* superfamilies, and clusters annotated  
13  
14 as LTR-retrotransposons by RepeatExplorer (see above) were then used as reference datasets for  
15  
16 further searches in order to identify previously unclassified elements using RepeatMasker, running  
17  
18 default parameters, but with -div set to 20.  
19  
20  
21  
22

23  
24 For determination of RE redundancy, approximately 32,000,000 raw *Potentilla* Illumina paired end  
25  
26 reads were randomly selected, corresponding to  $10.3\times$  genome coverage. After removal of organellar  
27  
28 contamination performed by mapping the reads to an in-house Rosaceae organellar database and the  
29  
30 removal of duplicate reads, a total of 25,206,510 reads corresponding to  $7.2\times$  equivalent genomic  
31  
32 coverage were used for redundancy analysis by mapping the reads to all REs characterized in the  
33  
34 *Potentilla* genome using CLC-BIO Genomic Workbench 8.0 (CLC-BIO, Aarhus, Denmark).  
35  
36 Mismatch cost, deletion cost, and insertion cost were fixed at 1, and similarity and length fraction  
37  
38 were both fixed at 0.9, 0.8, 0.5 or 0.4 to obtain high, medium, low, or very low stringencies,  
39  
40 respectively. As reads that mapped to multiple distinct sequences were few, and distributed randomly  
41  
42 throughout the dataset, the number of reads mapping to each RE was taken as the degree of  
43  
44 redundancy of that sequence within the genome. The effective abundance of a particular class of reads  
45  
46 was calculated as the proportion of the total number of reads mapped in each class, with respect to  
47  
48 the overall number of genomic reads mapped, using optimal stringency parameters, i.e. where further  
49  
50 relaxation of stringency did not significantly increase the number of mapped reads.  
51  
52  
53  
54  
55  
56

57  
58 The abundance of each single RE sequence in the genome was analysed by mapping *Potentilla* DNA  
59  
60 reads, corresponding to  $2\times$  genome coverage to the full-length REs characterised, one by one using  
61  
62  
63  
64  
65



1 BWA (alignment via Burrows–Wheeler transformation) version 0.7.5a-r405 (BWA,  
2 RRID:SCR\_010910) [56] running the following parameters: bwaaln -t 4 -l 12 -n 4 -k 2 -o 3 -e 3 -M  
3  
4 2 -O 6 -E 3. The resulting single-end mappings were resolved via the samse module of BWA, and the  
5  
6 output was converted to .bam file format using SAMtools version 0.1.19 [57]. Subsequently,  
7  
8 SAMtools was used to calculate the number of mapped reads for each alignment using the following  
9  
10 parameters: samtools view -c -F 4.  
11  
12  
13  
14  
15  
16

### 17 **Determination of RE insertion age**

18  
19 Retrotransposon insertion age was estimated through a sequence divergence comparison of the 5'-  
20  
21 and 3'-LTRs of each putative full-length retrotransposon. Synonymous substitution rates were  
22  
23 calculated for 50 pairs of orthologous genes of *P. micrantha* and *F. vesca*, using a time of divergence  
24  
25 of 24.22 million years [3]. Subsequently, the two LTRs were aligned with ClustalX software [56],  
26  
27 indels were eliminated, and the number of nucleotide substitutions was counted using DnaSP [57] for  
28  
29 each retrotransposon. The insertion times of retrotransposons with both LTRs were dated using the  
30  
31 Kimura two parameter (K2P) method [58], calculated using DnaSP, and a synonymous substitution  
32  
33 rate that is twofold that calculated for genes [16,17].  
34  
35  
36  
37  
38  
39  
40

### 41 **AVAILABILITY OF SUPPORTING DATA AND MATERIALS**

42  
43 The data set supporting the results of this article are available in the GenBank repository, project  
44  
45 number PRJEB18433. The genome reference sequence and gene predictions can be downloaded from  
46  
47 the GigaScience GigaDB repository [59].  
48  
49  
50  
51  
52

### 53 **FUNDING**

54  
55 This work was funded by a grant to the Fondazione Edmund Mach (FEM) from the Autonomous  
56  
57 Province of Trento grants office. A.C. acknowledges funding from the Department of Agriculture,  
58  
59 Food and Environment of Pisa University, Project 'Plantomics'.  
60  
61  
62  
63  
64  
65

1  
2 **CONFLICT OF INTERESTS**

3  
4  
5 The authors declare no competing interests.  
6  
7  
8

9  
10 **AUTHOR CONTRIBUTIONS**

11  
12 M.Buti performed the experiments, analysed and interpreted all data and authored the paper. M.M.,  
13  
14 P.S. and A.C. analysed sequence data and performed genome assemblies. K.E. and M. Brillì assisted  
15  
16 with experimental design, analysed and interpreted gene expression data and commented on and  
17  
18 contributed to the manuscript. L.N. and A.C. performed full-length retrotransposon isolation. E.B.,  
19  
20 F.M. and A.C. performed clustering, annotation and redundancy analyses of repetitive sequences.  
21  
22 E.B., F.M., L.N. and A.C. participated in the interpretation and discussion of results and contributed  
23  
24 to the writing of the paper. A.L and M.Borodovsky performed gene predictions and analysed and  
25  
26 interpreted the data. L.G., N.Š. assisted with experiments, interpreted data and contributed to the  
27  
28 manuscript. M.A. and J.W. assisted with genome assemblies and gene annotation. C.V. analysed and  
29  
30 interpreted phylogenetic data and contributed to the manuscript. R.V. commented on the manuscript.  
31  
32 D.J.S. designed the study, assisted with the experiments, analysed and interpreted the data and  
33  
34 authored the paper.  
35  
36  
37  
38  
39  
40  
41  
42

43  
44 **ADDITIONAL FILES**

45  
46 Additional File 1: Table S1. Illumina sequencing libraries used in the sequencing of the *Potentilla*  
47  
48 *micrantha* genome including fragment sizes and total genome depth of coverage.  
49

50  
51 Additional File 2: Table S2. PacBio RS sequencing kits and chemistries used for *Potentilla micrantha*  
52  
53 sequencing.  
54

55  
56 Additional File 3: Table S3. RNAseq read data used for gene prediction and number of splice sites  
57  
58 identified in the *Potentilla micrantha* genome.  
59

60  
61 Additional File 4: Table S4. *Potentilla micrantha* and *Fragaria vesca* RNAseq reads statistics.  
62

1 Additional File 5: Fig S1. Distribution of predicted genes *Potentilla micrantha* and *Fragaria vesca*  
2 mapped, blasted and GO-annotated by BLAST2GO analysis.

3  
4 Additional File 6: Fig S2. The differential gene expression profiles between the four developmental  
5 stages of fruit development studied in *F. vesca* and *P. micrantha*.

6  
7 Additional File 7: Fig S3. The overall abundance of different classes of transposons within the  
8  
9  
10 *Potentilla micrantha* genome according to the analyses performed using RepeatExplorer.

11  
12 Additional File 8: Fig S4. Genome proportion in *Potentilla micrantha* and *Fragaria vesca* of 291  
13 repeats clustered using RepeatExplorer. Other repeats include satellite DNAs, pararetroviruses, and  
14  
15 one LINE.  
16  
17  
18  
19  
20  
21  
22

## 23 24 REFERENCES

- 25  
26 1. Eriksson T, Donoghue MJ, Hibbs MS. Phylogenetic analysis of *Potentilla* using DNA sequences  
27 of nuclear ribosomal internal transcribed spacers (ITS), and implications for the classification of  
28  
29 Rosoideae (Rosaceae). *Plant Syst. Evol.* 1998; 211:155–79.  
30  
31
- 32 2. Potter D, Eriksson T, Evans RC et al. Phylogeny and classification of Rosaceae. *Plant Syst. Evol.*  
33  
34 2007; 266:5–43.  
35  
36
- 37 3. Njuguna W, Liston A, Cronn R, Ashman T-L, Bassil N. Insights into phylogeny, sex function  
38  
39 and age of *Fragaria* based on whole chloroplast genome sequencing. *Mol. Phylogenet. Evol.* 2013;  
40  
41 66:17–29.  
42  
43
- 44 4. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome  
45  
46 of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 2011; 43:109–16.  
47  
48
- 49 5. Jung S, Cestaro A, Troggio M, Main D, Zheng P, Cho I, et al. Whole genome comparisons of  
50  
51 *Fragaria*, *Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies.  
52  
53 *BMC Genomics* 2012; 13:129.  
54  
55
- 56 6. Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu S-H, et al. Comparative  
57  
58 transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.*  
59  
60  
61

- 1 2012; 71:492–502.
- 2 7. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, et al. A genome  
3  
4  
5 3 triplication associated with early diversification of the core eudicots. *Genome Biol.* 2012; 13:R3.  
6
- 7 4 8. Ferrarini M, Moretto M, Ward JA, Šurbanovski N, Stevanović V, Giongo L, et al. An evaluation  
8  
9  
10 5 of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC*  
11  
12 6 *Genomics* [Internet]. BioMed Central; 2013 [cited 2016 Aug 8];14:670. Available from:  
13  
14 7 <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-670>  
15  
16
- 17 8 9. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Letter to the editor. *Cytometry* [Internet]. Wiley  
18  
19 9 Subscription Services, Inc., A Wiley Company; 2003 [cited 2016 Aug 9];51A:127–8. Available  
20  
21  
22 10 from: <http://doi.wiley.com/10.1002/cyto.a.10013>  
23
- 24 11 10. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading  
25  
26  
27 12 Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. Liu Z, editor. *PLoS*  
28  
29 13 *One* [Internet]. Public Library of Science; 2012 [cited 2016 Aug 8];7:e47768. Available from:  
30  
31 14 <http://dx.plos.org/10.1371/journal.pone.0047768>  
32  
33
- 34 15 11. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: assessing  
35  
36 16 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*  
37  
38  
39 17 [Internet]. 2015 [cited 2017 Nov 2];31:3210–2. Available from:  
40  
41 18 <http://www.ncbi.nlm.nih.gov/pubmed/26059717>  
42  
43
- 44 19 12. Kang C, Darwish O, Geretz A, Shahan R, Alkharouf N, Liu Z. Genome-Scale Transcriptomic  
45  
46 20 Insights into Early-Stage Fruit Development in Woodland Strawberry *Fragaria vesca*. *Plant Cell*  
47  
48  
49 21 [Internet]. 2013;25:1960–78. Available from:  
50
- 51 22 13. Day RC, Herridge RP, Ambrose BA, Macknight RC. Transcriptome Analysis of Proliferating  
52  
53 23 *Arabidopsis* Endosperm Reveals Biological Implications for the Control of Syncytial Division,  
54  
55  
56 24 Cytokinin Signaling, and Gene Expression Regulation. *PLANT Physiol.* [Internet]. American  
57  
58 25 Society of Plant Biologists; 2008 [cited 2016 Aug 10];148:1964–84. Available from:  
59  
60  
61 26 <http://www.plantphysiol.org/cgi/doi/10.1104/pp.108.128108>  
62  
63  
64  
65

- 1 14. Hehenberger E, Kradolfer D, Köhler C. Endosperm cellularization defines an important  
1  
2 2 developmental transition for embryo development. *Development* [Internet]. 2012 [cited 2016 Aug  
3  
4 3 10];139:2031–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22535409>  
5  
6  
7 4 15. Fang S-C, Fernandez DE. Effect of regulated overexpression of the MADS domain factor  
8  
9 5 AGL15 on flower senescence and fruit maturation. *Plant Physiol.* [Internet]. 2002 [cited 2016 Aug  
10  
11  
12 6 10];130:78–89. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12226488>  
13  
14 7 16. Sanmiguel P, Bennetzen JL. Evidence that a Recent Increase in Maize Genome Size was  
15  
16  
17 8 Caused by the Massive Amplification of Intergene Retrotransposons. *Ann. Bot. Oxford University*  
18  
19 9 Press; 1998;82:37–44.  
20  
21  
22 10 17. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl.*  
23  
24 11 Acad. Sci. U. S. A. [Internet]. National Academy of Sciences; 2004 [cited 2016 Aug 9];101:12404–  
25  
26  
27 12 10. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15240870>  
28  
29 13 18. VanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, et al. The genome of  
30  
31 14 black raspberry (*Rubus occidentalis*). *Plant J.* [Internet]. 2016 [cited 2016 Aug 16]; Available from:  
32  
33  
34 15 <http://www.ncbi.nlm.nih.gov/pubmed/27228578>  
35  
36 16 19. Dickson EE, Arumuganathan K, Kresovich S, Doyle JJ, Kresovich S, Doyle2 JJ. Nuclear DNA  
37  
38  
39 17 Content Variation within the Rosaceae NUCLEAR DNA CONTENT VARIATION WITHIN THE  
40  
41 18 ROSACEAE'. *Am. J. Bot. Am. J. Bot. Am. J. Bot.* [Internet]. 1992 [cited 2016 Nov 5];79:1081–6.  
42  
43  
44 19 Available from: [http://scholarcommons.sc.edu/biol\\_facpub](http://scholarcommons.sc.edu/biol_facpub)  
45  
46 20 20. Meng R, Finn C. Determining Ploidy Level and Nuclear DNA Content in *Rubus* by Flow  
47  
48  
49 21 Cytometry. *J. Am. Soc. Hortic. Sci. American Society for Horticultural Science*; 2002;127:767–75.  
50  
51 22 21. Rajapakse S, Byrne DH, Zhang L, Anderson N, Arumuganathan K, Ballard RE. Two genetic  
52  
53  
54 23 linkage maps of tetraploid roses. *TAG Theor. Appl. Genet.* [Internet]. Springer-Verlag; 2001 [cited  
55  
56 24 2016 Nov 5];103:575–83. Available from: <http://link.springer.com/10.1007/PL00002912>  
57  
58 25 22. Yokoya K, Roberts A V., Mottley J, Lewis R, Brandham PE. Nuclear DNA Amounts in Roses.  
59  
60  
61 26 *Ann. Bot.* [Internet]. Oxford University Press; 2000 [cited 2016 Nov 5];85:557–61. Available from:

- 1 <http://aob.oxfordjournals.org/cgi/doi/10.1006/anbo.1999.1102>
- 2 23. Vitte C, Fustier M-A, Alix K, Tenaillon MI. The bright side of transposons in crop evolution.  
3  
4 Brief. Funct. Genomics [Internet]. Oxford University Press; 2014 [cited 2016 Aug 15];13:276–95.  
5  
6 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24681749>
- 7 24. Suzuki M, Kamide Y, Nagata N, Seki H, Ohyama K, Kato H, et al. Loss of function of 3-  
8  
9 hydroxy-3-methylglutaryl coenzyme A reductase 1 (HMG1) in Arabidopsis leads to dwarfing, early  
10  
11 senescence and male sterility, and reduced sterol levels. Plant J. [Internet]. 2004 [cited 2017 Nov  
12  
13 2];37:750–61. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/14871314>
- 14 25. Schrick K, Debolt S, Bulone V. Deciphering the molecular functions of sterols in cellulose  
15  
16 biosynthesis. Front. Plant Sci. [Internet]. Frontiers Media SA; 2012 [cited 2017 Nov 2];3:84.  
17  
18 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22639668>
- 19 26. Smaczniak C, Immink RGH, Angenent GC, Kaufmann K, Adamczyk BJ, Fernandez DE, et al.  
20  
21 Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent  
22  
23 studies. Development [Internet]. Oxford University Press for The Company of Biologists Limited;  
24  
25 2012 [cited 2016 Aug 15];139:3081–98. Available from:  
26  
27 <http://www.ncbi.nlm.nih.gov/pubmed/22872082>
- 28 27. Shirzadi R, Andersen ED, Bjerkan KN, Gloeckle BM, Heese M, Ungru A, et al. Genome-wide  
29  
30 transcript profiling of endosperm without paternal contribution identifies parent-of-origin-  
31  
32 dependent regulation of AGAMOUS-LIKE36. PLoS Genet. [Internet]. 2011 [cited 2016 Aug  
33  
34 16];7:e1001303. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21379330>
- 35 28. Harding EW, Tang W, Nichols KW, Fernandez DE, Perry SE. Expression and maintenance of  
36  
37 embryogenic potential is enhanced through constitutive expression of AGAMOUS-Like 15. Plant  
38  
39 Physiol. [Internet]. 2003 [cited 2016 Aug 16];133:653–63. Available from:  
40  
41 <http://www.ncbi.nlm.nih.gov/pubmed/14512519>
- 42 29. Serivichyaswat P, Ryu H-S, Kim W, Kim S, Chung KS, Kim JJ, et al. Expression of the floral  
43  
44 repressor miRNA156 is positively regulated by the AGAMOUS-like proteins AGL15 and AGL18.  
45  
46

- 1 Mol. Cells [Internet]. Korean Society for Molecular and Cellular Biology; 2015 [cited 2016 Aug  
2 16];38:259–66. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25666346>  
3
- 4 30. Chen D-H, Ronald PC. A Rapid DNA Minipreparation Method Suitable for AFLP and Other  
5 PCR Applications. *Plant Mol. Biol. Report.* [Internet]. Kluwer Academic Publishers; 1999 [cited  
6 2016 Aug 8];17:53–7. Available from: <http://link.springer.com/10.1023/A:1007585532036>  
7
- 8 31. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS:  
9 de novo assembly of whole-genome shotgun microreads. *Genome Res.* [Internet]. 2008 [cited 2016  
10 Aug 8];18:810–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18340039>  
11
- 12 32. Smit AFA, Hubley R. RepeatModeler - 1.0.7 [Internet]. 2013. Available from:  
13 <http://www.repeatmasker.org/RepeatModeler.html>  
14
- 15 33. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. 2013. Available from:  
16 <http://www.repeatmasker.org/>  
17
- 18 34. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic  
19 training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* [Internet]. Oxford University  
20 Press; 2014 [cited 2016 Aug 8];42:e119. Available from:  
21 <http://www.ncbi.nlm.nih.gov/pubmed/24990371>  
22
- 23 35. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast  
24 universal RNA-seq aligner. *Bioinformatics* [Internet]. Oxford University Press; 2013 [cited 2016  
25 Aug 8];29:15–21. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23104886>  
26
- 27 36. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics.  
28 *Int. J. Plant Genomics* [Internet]. Hindawi Publishing Corporation; 2008 [cited 2016 Aug  
29 8];2008:619832. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18483572>  
30
- 31 37. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. InParanoid 7: new  
32 algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* [Internet]. 2010 [cited  
33 2016 Aug 10];38:D196-203. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19892828>  
34
- 35 38. Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, et al. GDR (Genome Database for  
36

- 1 Rosaceae): integrated web-database for Rosaceae genomics and genetics data. *Nucleic Acids Res.*  
2 [Internet]. Oxford University Press; 2008 [cited 2016 Aug 9];36:D1034-40. Available from:  
3  
4  
5 3 <http://www.ncbi.nlm.nih.gov/pubmed/17932055>  
6  
7 4 39. QuickParanoid - A tool for ortholog clustering, <http://pl.postech.ac.kr/QuickParanoid/> Accessed  
8  
9 5 30 July 2016.  
10  
11  
12 6 40. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput  
13  
14 7 Sequence Data [Internet]. 2010. Available from:  
15  
16  
17 8 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>  
18  
19 9 41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.  
20  
21  
22 10 Bioinformatics [Internet]. Oxford University Press; 2014 [cited 2016 Aug 9];30:2114–20. Available  
23  
24 11 from: <http://www.ncbi.nlm.nih.gov/pubmed/24695404>  
25  
26  
27 12 42. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* [Internet].  
28  
29 13 NIH Public Access; 2012 [cited 2016 Aug 8];9:357–9. Available from:  
30  
31  
32 14 <http://www.ncbi.nlm.nih.gov/pubmed/22388286>  
33  
34 15 43. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying  
35  
36 16 mammalian transcriptomes by RNA-Seq. *Nat. Methods* [Internet]. Nature Publishing Group; 2008  
37  
38  
39 17 [cited 2016 Aug 8];5:621–8. Available from: <http://www.nature.com/doifinder/10.1038/nmeth.1226>  
40  
41 18 44. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J.*  
42  
43 19 *Plant Genomics* [Internet]. Hindawi Publishing Corporation; 2008 [cited 2016 Aug 8];2008:619832.  
44  
45  
46 20 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18483572>  
47  
48  
49 21 45. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software  
50  
51 22 environment for integrated models of biomolecular interaction networks. *Genome Res.* [Internet]. Cold  
52  
53 23 Spring Harbor Laboratory Press; 2003 [cited 2017 Nov 3];13:2498–504. Available from:  
54  
55 24 <http://www.ncbi.nlm.nih.gov/pubmed/14597658>  
56  
57  
58 25 46. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD:  
59  
60 26 NCBI’s conserved domain database. *Nucleic Acids Res.* [Internet]. 2015 [cited 2016 Aug  
61  
62  
63  
64  
65



- 1 8];43:D222-6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25414356>
- 2 47. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.  
3  
4  
5 3 Bioinformatics [Internet]. 2010 [cited 2016 Aug 8];26:841–2. Available from:  
6  
7 4 <http://www.ncbi.nlm.nih.gov/pubmed/20110278>  
8
- 9 48. Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, et al. SATE-II: very fast and  
10 5 accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Syst.  
11  
12 6 Biol. [Internet]. Oxford University Press; 2012 [cited 2016 Aug 9];61:90–106. Available from:  
13  
14 7 <http://www.ncbi.nlm.nih.gov/pubmed/22139466>  
15  
16  
17 8
- 19 9 49. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version  
20  
21 7.0 for Bigger Datasets. Mol. Biol. Evol. [Internet]. 2016;33:1870–4. Available from:  
22 10  
23  
24 11 <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw054>  
25  
26  
27 12
- 29 13 50. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic  
28  
29 13 trees made easy. Nucleic Acids Res. [Internet]. Oxford University Press; 2011 [cited 2016 Aug  
30  
31 14  
32 14 8];39:W475-8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21470960>  
33
- 34 15 51. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, et al. The  
35  
36 16 Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information  
37  
38  
39 17 retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res. [Internet]. Oxford  
40  
41 18 University Press; 2001 [cited 2016 Aug 8];29:102–5. Available from:  
42  
43  
44 19 <http://www.ncbi.nlm.nih.gov/pubmed/11125061>  
45
- 46 20 52. Novák P, Neumann P, Pech J, Steinhaisl J, Macas J. RepeatExplorer: a Galaxy-based web  
47  
48  
49 21 server for genome-wide characterization of eukaryotic repetitive elements from next-generation  
50  
51 22 sequence reads. Bioinformatics [Internet]. 2013 [cited 2016 Aug 9];29:792–3. Available from:  
52  
53  
54 23 <http://www.ncbi.nlm.nih.gov/pubmed/23376349>  
55
- 56 24 53. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update,  
57  
58 25 a database of eukaryotic repetitive elements. Cytogenet. Genome Res. [Internet]. Karger Publishers;  
59  
60  
61 26 2005 [cited 2016 Aug 9];110:462–7. Available from:

1 <http://www.karger.com/?doi=10.1159/000084979>

2 54. Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR  
3 retrotransposons. *Nucleic Acids Res.* [Internet]. Oxford University Press; 2007 [cited 2016 Aug  
4 8];35:W265-8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17485477>

5 55. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for  
6 genomic DNA and protein sequence analysis. *Gene* [Internet]. 1995 [cited 2016 Aug 8];167:GC1-  
7 10. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8566757>

8 56. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.  
9 *Bioinformatics* [Internet]. 2009 [cited 2016 Aug 9];25:1754–60. Available from:  
10 <http://www.ncbi.nlm.nih.gov/pubmed/19451168>

11 57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence  
12 Alignment/Map format and SAMtools. *Bioinformatics* [Internet]. 2009 [cited 2016 Aug  
13 9];25:2078–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19505943>

14 58. Kimura M. A simple method for estimating evolutionary rates of base substitutions through  
15 comparative studies of nucleotide sequences. *J. Mol. Evol.* [Internet]. 1980 [cited 2016 Aug  
16 9];16:111–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/7463489>

17 59. Buti M, Moretto M, Barghini E, Mascagni F, Natali N, Brilli M, et al (2018). Supporting data  
18 for ‘The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to  
19 *Fragaria vesca* (the woodland strawberry)’. *GigaScience Database* 2018.  
20 <http://dx.doi.org/10.5524/100407>

## 21 **FIGURE LEGENDS AND TABLES**

22 **Figure 1.** Comparison of *Fragaria vesca* and *Potentilla micrantha* morphology for leaves, flowers  
23 and fruits.

24 **Figure 2a.** Anchoring of five *Potentilla micrantha* genome scaffolds to the *Fragaria vesca* Fvb  
25 pseudomolecules *Fvb2* and *Fvb4* demonstrating the microsynteny between the *F. vesca* and  
26

1 *P. micrantha* genomes (numbers in parentheses below the scaffold names indicate the number of  
2 genes contained in each split syntenic block.  
3  
4 **Figure 2b.** A comparison of the seven pseudomolecules of the *F. vesca* genome with eight  
5  
6  
7 *P. micrantha* sequencing scaffolds, highlighting the major translocation events identified between the  
8  
9  
10 two species in this investigation.  
11  
12 **Figure 3.** *Potentilla micrantha* flower/fruit developmental stages used for RNA extraction.  
13  
14 **Figure 4.** Differentially expressed genes during fruit development in *Potentilla micrantha* and  
15  
16  
17 *Fragaria vesca*. Volcano plots of differential expression analysis between the four developmental  
18  
19 stages A-B-C-D in *P. micrantha* and *F. vesca*. Using a cut-off of  $\sqrt{\text{MSR}} > 2.00$  and  $p\text{-value} < 10^{-3}$ ,  
20  
21 1,556 genes were differentially expressed in *P. micrantha*, whilst 816 genes were differentially  
22  
23 expressed in *F. vesca*.  
24  
25  
26 **Figure 5.** Over-represented GO-slim categories in *Fragaria vesca* and *Potentilla micrantha* DEGs  
27  
28  
29 sets. The circles are shaded based on significance level (yellow = FDR below 0.05), and the radius of  
30  
31 each circle is proportional to the number of genes included in each GO-slim category.  
32  
33  
34 **Figure 6.** Heatmap comparing the log expression values of 205 genes (orthologs of both *Fragaria*  
35  
36 *vesca* and *Potentilla micrantha*) The rows (genes) were sorted using hierarchical clustering using  
37  
38 'correlation' distance and 'complete' linkage. A-D correspond to the four developmental stages defined  
39  
40 in the methods section.  
41  
42  
43 **Figure 7.** A Maximum Likelihood-based phylogenetic reconstruction of the *Potentilla micrantha* and  
44  
45  
46 *Fragaria vesca* genes containing MADS-box motifs, along with the relative gene expression levels  
47  
48 for each gene. Categories A-D refer to the developmental stages defined in the methods. Filled circles  
49  
50 represent the relative level of support for each relationship defined in the Maximum Likelihood  
51  
52 analysis.  
53  
54  
55 **Figure 8.** The three identified clades of orthologous MADS-box motif containing genes that were not  
56  
57 expressed or poorly expressed in *Potentilla micrantha* but highly expressed in *Fragaria vesca*.  
58  
59 Categories A-D refer to the four developmental stages defined in the methods.  
60  
61  
62  
63  
64  
65

1  
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

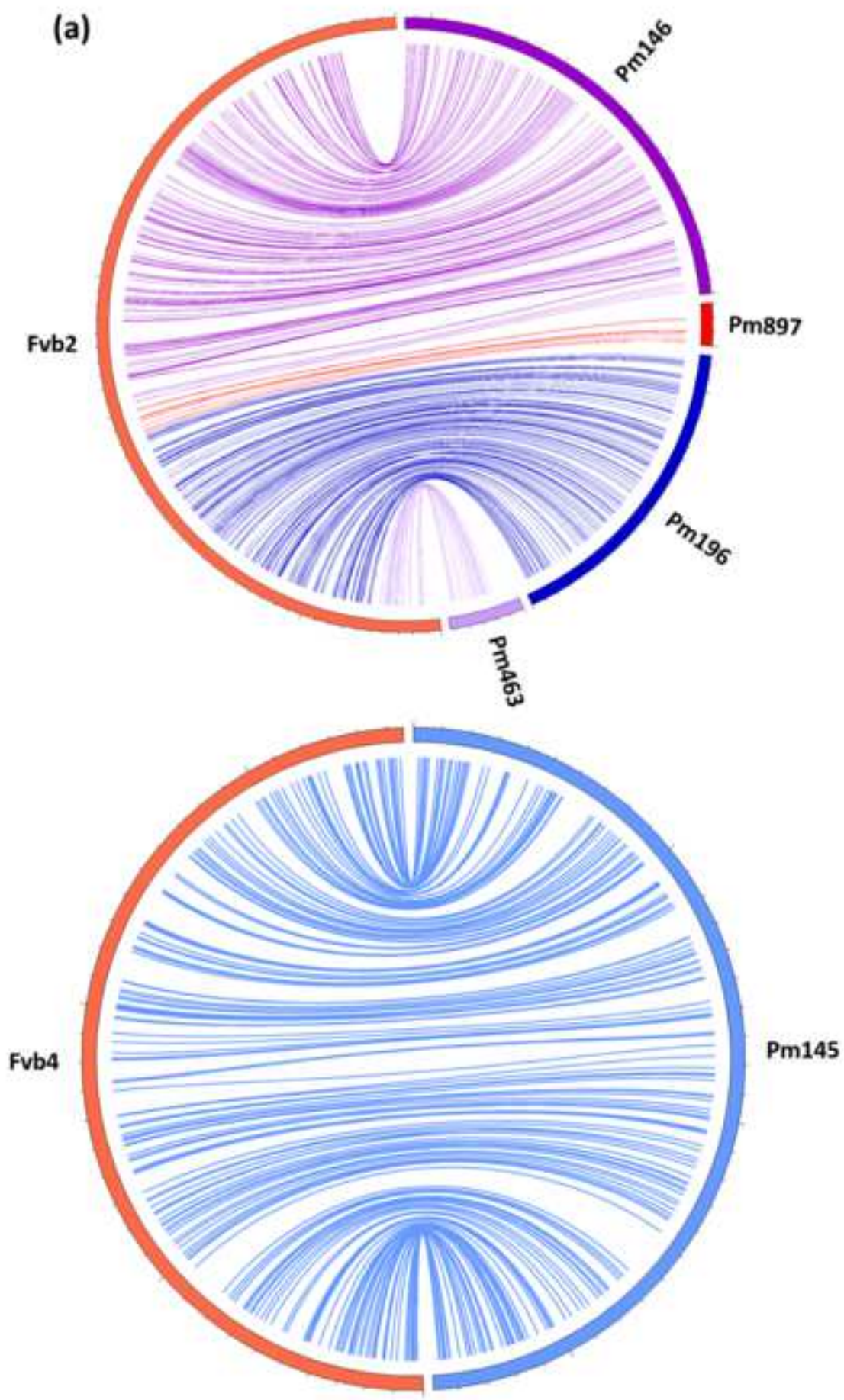
**Table 1.** *Potentilla micrantha* assembly stats

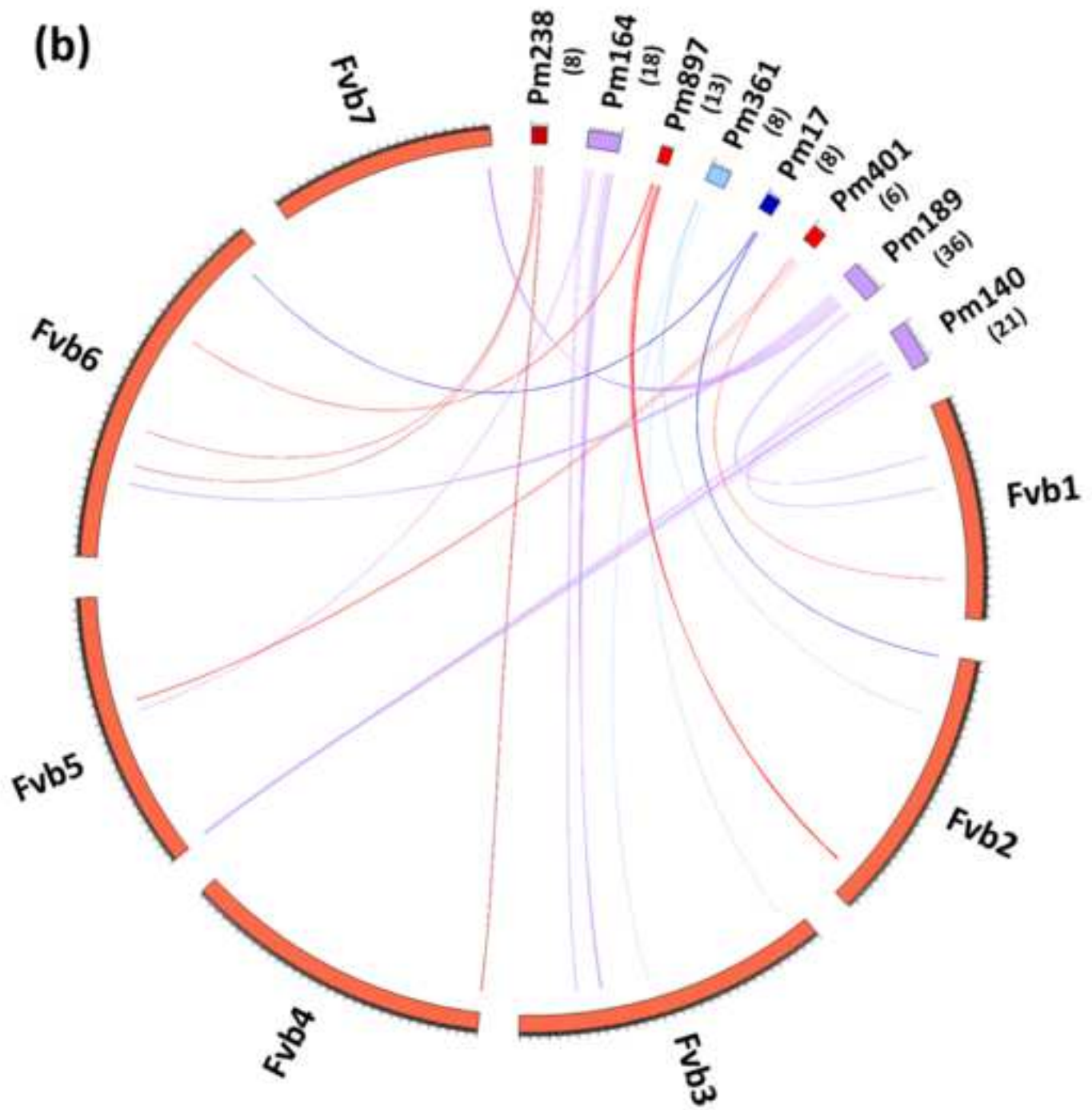
	ALLPATHS-LG Illumina data	PacBio PBJelly
Number of scaffolds	2,866	2,674 (-6.7%)
Total size of scaffolds	315,266,043	326,533,584 (+3.5%)
Longest scaffold	3,162,838	3,488,351 (+9.3%)
N50 scaffold length	318,490	335,712 (+5.1%)
Gapped Ns in scaffolds	67,706,454	27,311,787 (-59.7%)
Number of contigs	33,026	n/a
Number of contigs in scaffolds	32,063	n/a
Total size of contigs	247,565,733	n/a
N50 contig length	16,235	n/a

1 **Table 2.** Annotation of 505 full-length LTR-retrotransposons of *Potentilla micrantha*.

2	3	4	5	6
Superfamily	Family	Number	Percentage	
7	<i>Ty1-Copia</i>	<i>AleI/Retrofit</i>	14	2.77
8		<i>AleII</i>	26	5.15
9		<i>Angela</i>	20	3.96
10		<i>Bianca</i>	114	22.57
11		<i>Ivana</i>	23	4.55
12		<i>Maximus/SIRE</i>	10	1.98
13		<i>TAR/Tork</i>	11	2.18
14		Unknown	2	0.40
15		Total	220	43.56
16	<i>Ty3-Gypsy</i>	<i>Athila</i>	3	0.59
17		<i>Chromovirus</i>	42	8.32
18		<i>Ogre/TAT</i>	186	36.83
19		Unknown	25	4.95
20		Total	256	50.69
21	Unclassified		29	5.74

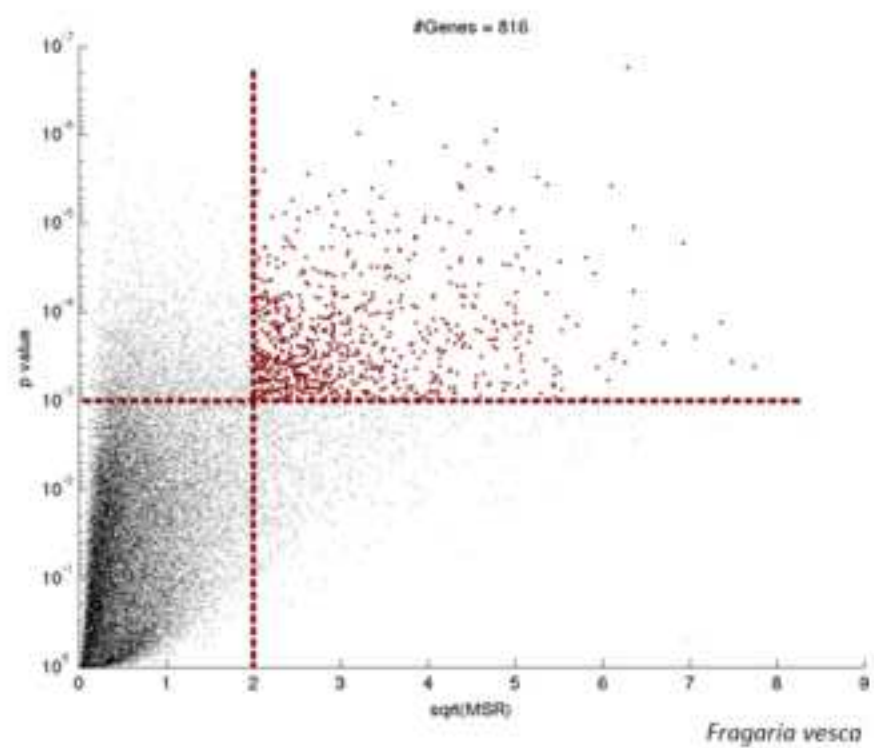
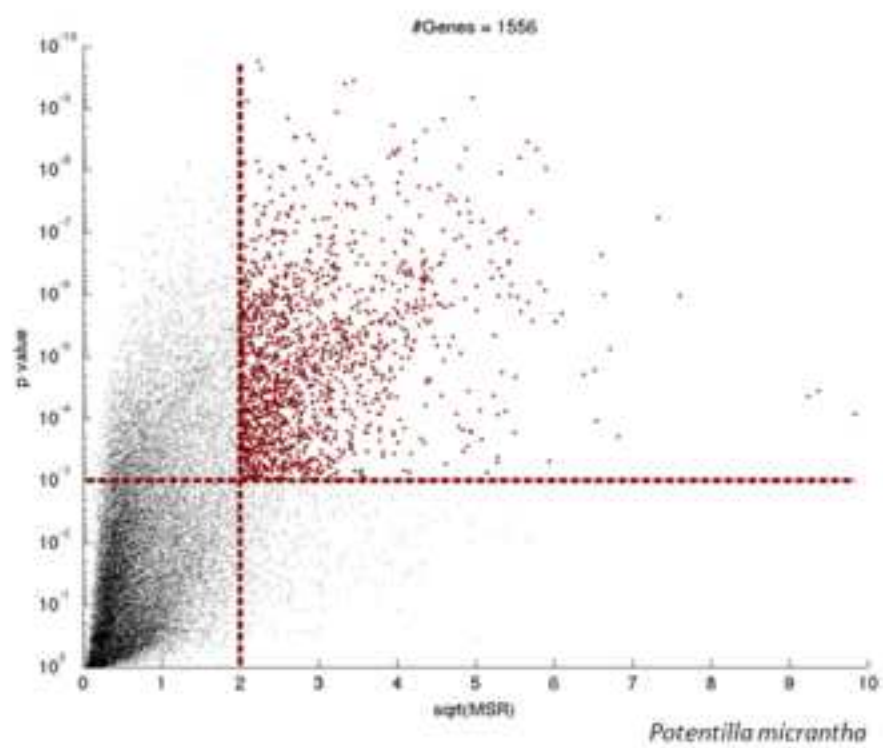


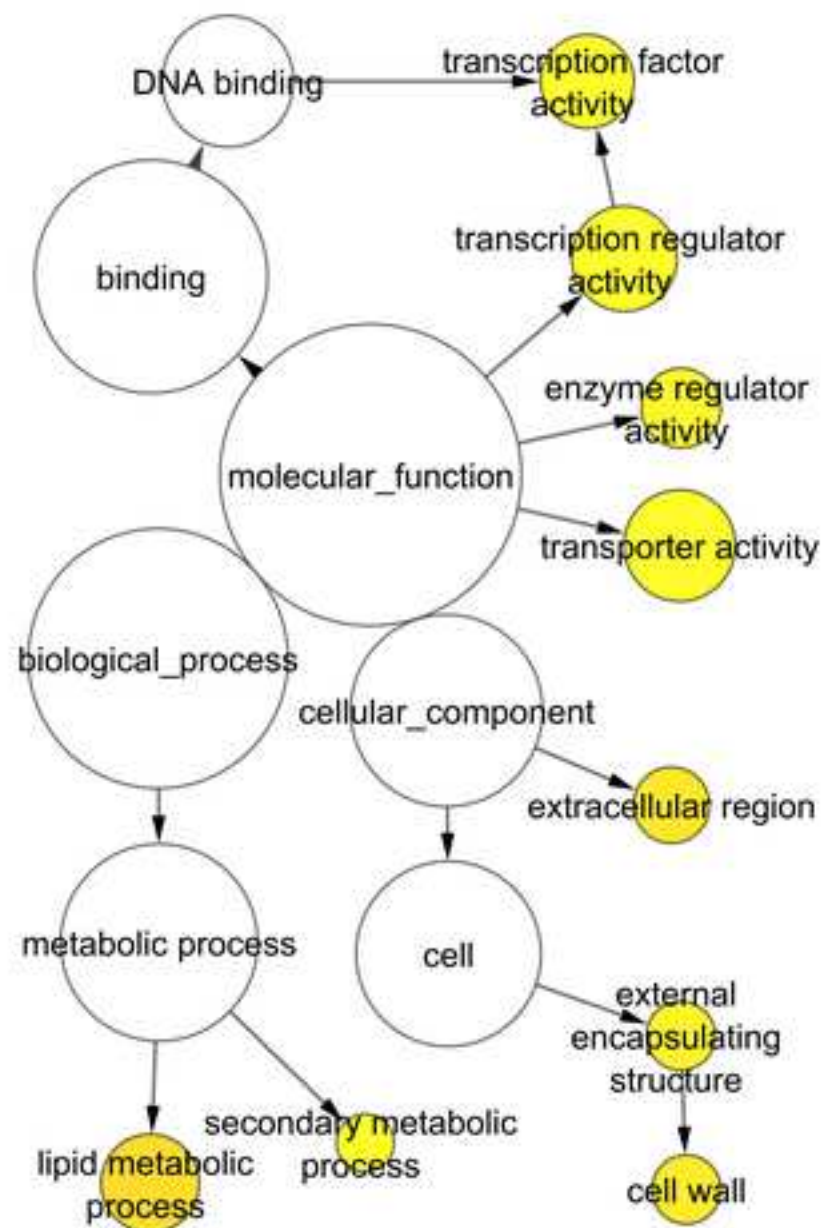
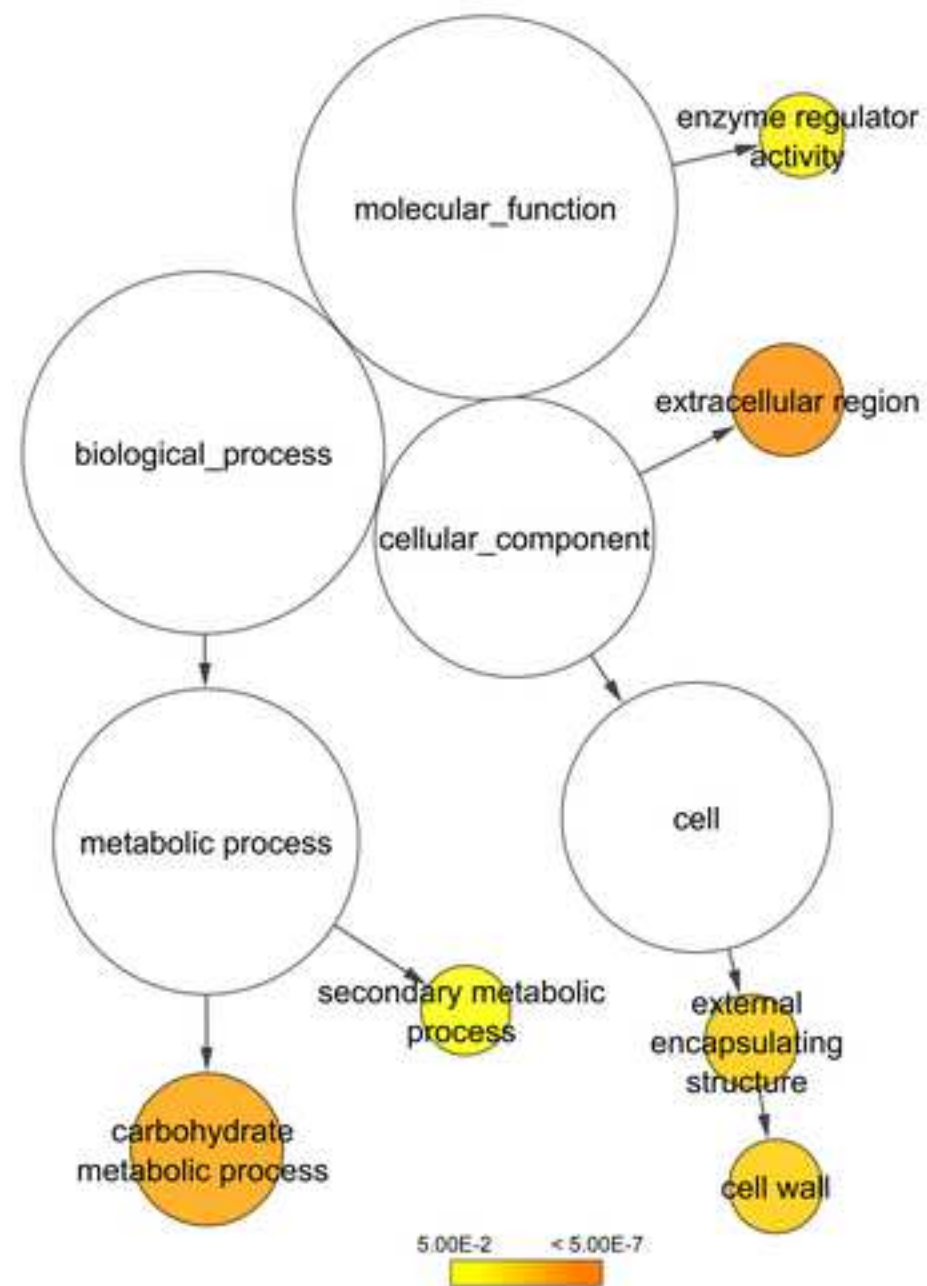










*Fragaria vesca**Potentilla micrantha*

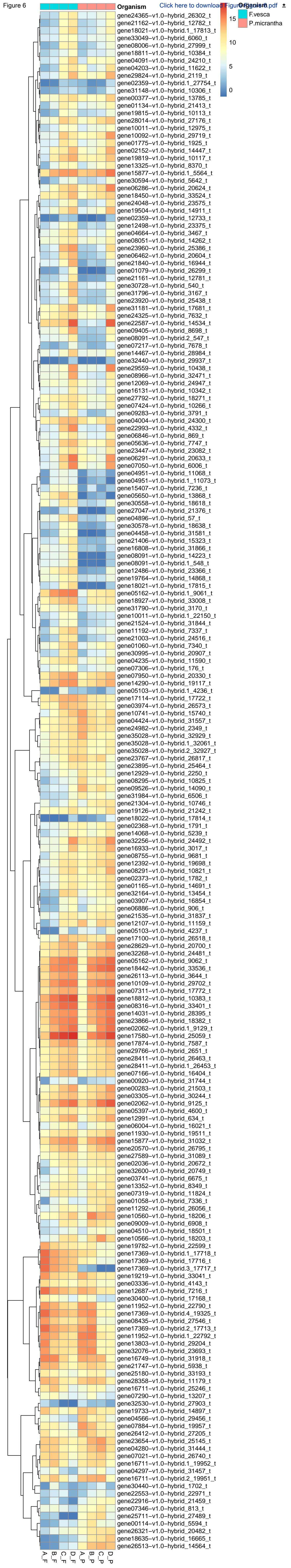
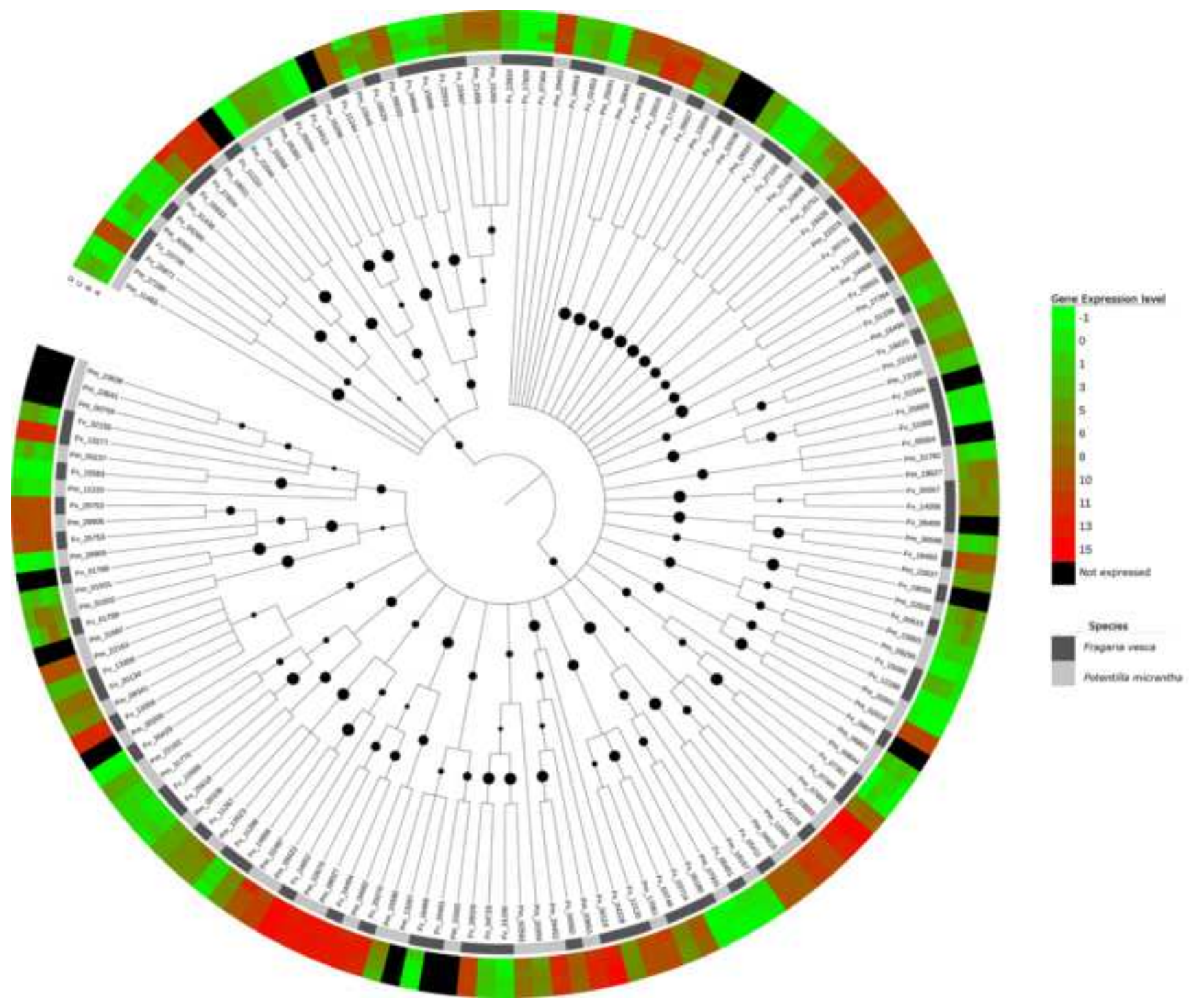
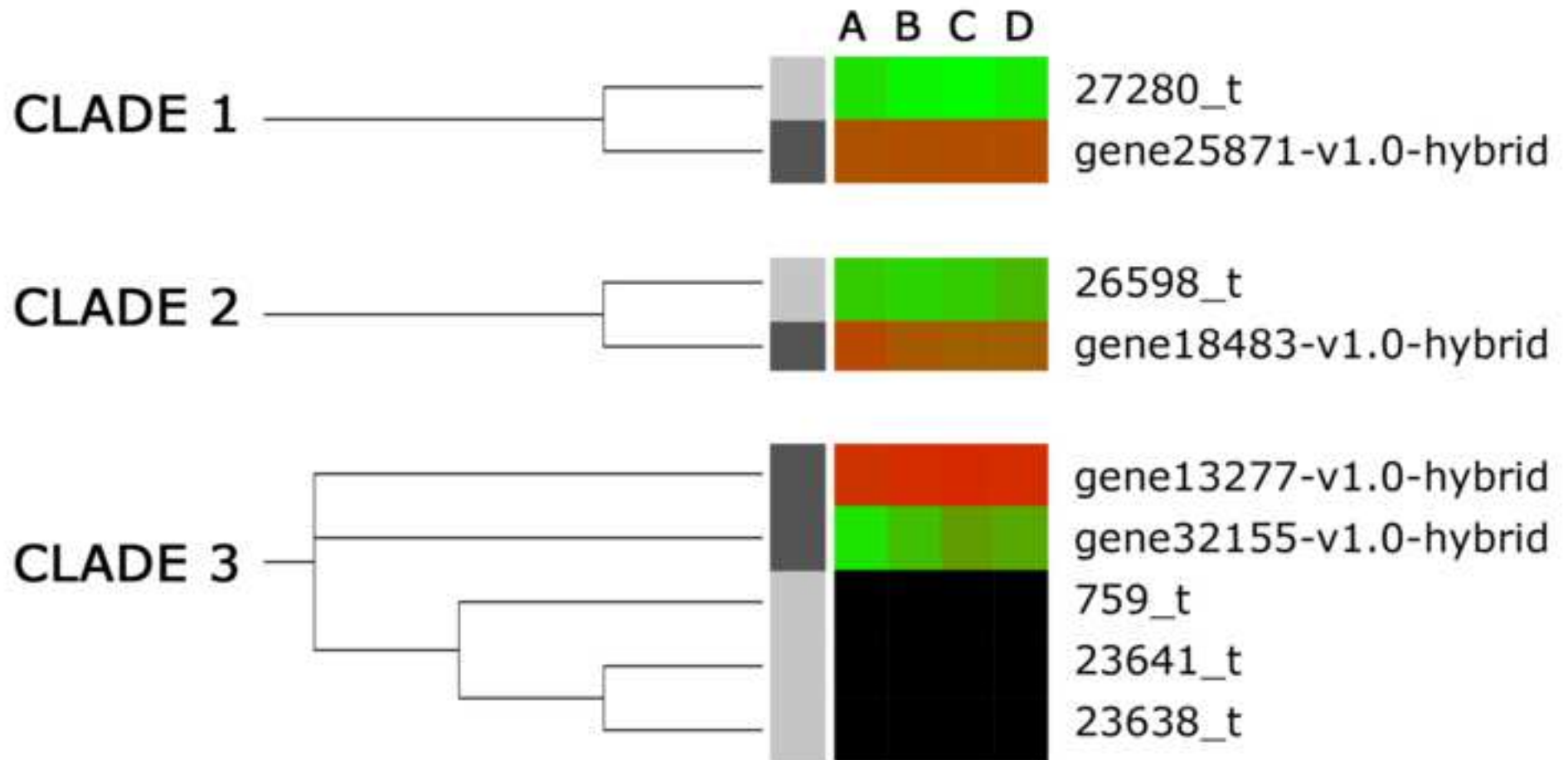


Figure 7

[Click here to download Figure Figure 7.tif](#)







Click here to access/download  
**Supplementary Material**  
Additional\_File\_1\_Table S1.docx





Click here to access/download  
**Supplementary Material**  
Additional\_File\_2\_Table S2.docx







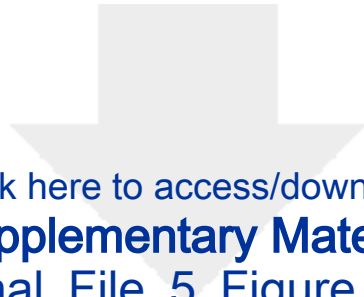
Click here to access/download  
**Supplementary Material**  
Additional\_File\_3\_Table S3.docx





Click here to access/download  
**Supplementary Material**  
Additional\_File\_4\_Table\_S4.docx





Click here to access/download  
**Supplementary Material**  
Additional\_File\_5\_Figure S1.docx





Click here to access/download  
**Supplementary Material**  
Additional\_File\_6\_Figure\_S2.png





Click here to access/download  
**Supplementary Material**  
Additional\_File\_7\_Figure\_S3.tif





Click here to access/download  
**Supplementary Material**  
Additional\_File\_8\_Figure\_S4.docx

