

# GigaScience

## 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-17-00155R2
<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>	Matteo Buti, PhD Marco Moretto, PhD Elena Barghini, PhD Flavia Mascagni, PhD Lucia Natali, PhD Matteo Brilli, PhD Alexandre Lomsadze, PhD Paolo Sonogo, PhD Lara Giongo, PhD Michael Alonge, MSc

	Riccardo Velasco, PhD
	Claudio Varotto, PhD
	Nada Surbanovski, PhD
	Mark Borodovsky, PhD
	Judson A Ward, PhD
	Kristoff Engelen, PhD
	Alessandro Cestaro, PhD
	Andrea Cavallini, PhD
	Daniel James Sargent, PhD
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	<p>Dr Daniel James Sargent  Driscoll's Genetics Limited  East Malling Enterprise Centre  New Road  East Malling  Kent, ME19 6BJ, UK  4th January 2018</p> <p>Dear Dr Zauner,</p> <p>Please find enclosed our revised manuscript: GIGA-D-17-00155R2 The genome sequence and transcriptome of <i>Potentilla micrantha</i> and their comparison to <i>Fragaria vesca</i> (the woodland strawberry), which has been reformatted in the style of a Data Note.</p> <p>Below we provide a point-by-point description of the changes made in response to the reviewers comments and hope that the changes we have made will be sufficient for the paper to be acceptable for publication.</p> <p>We look forward to hearing from you in due course regarding our revised paper.</p> <p>Very best regards,</p> <p>Dan Sargent (on behalf of all authors).</p> <p>Responses to Reviewer's comments:</p> <p>Q1. Overall, I am pleased that the reviewers agree that the latest, revised version of your manuscript presents useful and valuable data. However, I also agree with reviewer 1 that, in the absence of more in-depth analyses, the main value of the paper is now indeed the presentation of a high-quality resource, rather than new biological insights.</p> <p>R1. Many thanks for the useful comments. We agree with your feeling that the paper describes more of a technical advance than a description of new biological insights at this stage, but hope we have provided a valuable dataset that can be used for future biological investigations.</p> <p>Q2. I therefore feel that your revised manuscript may be more suitable for the format of a "Data Note". Data notes are indexed in the same way as research articles, but the emphasis is on presenting an exceptional resource, together with description and validation of the dataset. If you agree that we consider your next revised version as a data note, please adjust the format (see: <a href="https://academic.oup.com/gigascience/pages/data_note">https://academic.oup.com/gigascience/pages/data_note</a>). I feel that this would require only minor adjustments to the text itself (e.g. the discussion part could be revised to demonstrate validation and re-use cases of the data). You can keep the Background section for the Data Note format (this is not quite clear from our Instructions).</p>

Please let me know if you agree to this suggestion, and I'm happy to answer any questions regarding the Data Note format.

R2. We are happy for our paper to be published as a data note and have revised the manuscript accordingly (following the guidelines and taking recently published data note articles as examples).

Reviewer reports:

Q3. Reviewer #1: The authors addressed some of the main issues appropriately (synteny, figures etc.). However, for other issues like the transposon section, the hormonal treatment analysis and the miR1511 analysis the main action was just to shorten or completely drop the part. This is Ok and/or was suggested but on the other hand, no real efforts were made to strengthen the comparison/fleshy fruit analyses (or any other analytical part) and most of my suggestions/questions for this and also the annotation and gene expression part were simply ignored. As a result, this study as it stands now mostly provides an "extended description" of the resources generated (although potentially valuable) with clear shortcomings in the analysis and interpretation of the data. Along that lines, I appreciate that the authors in the new version resign from claiming analytical results not there or possible.

R3. We appreciate the reviewer's useful and constructive criticisms on the previous version of our paper and whilst we tried to incorporate as many of the suggested changes as was possible, the lead authors did not have the resources available to make all the suggested improvements, and as such we opted to remove the unsupported claims and refocus the paper as a technical report.

Additional comments:

Q4. a.) Your BUSCO analysis shows that you are missing ~6% of the BUSCO genes from the genome sequence (present) to the final gene prediction (absent). Are they completely missing in the gene prediction or fragmented etc.?

R4. We have provided more detail in the manuscript, detailing the complete and fragmented BUSCO sequences retrieved. We have also performed a BUSCO analysis on the *F. vesca* gene predictions and provided a comparison of the two to demonstrate that the *P. micrantha* set is virtually as complete as the *F. vesca* set.

Q5. b.) I still wonder about the ~9,000 gene predictions not showing a hit on the *F. vesca* pseudomolecules...do those genes have functional annotation and expression support?

R5. Our sincere apologies for not making this section clearer (it was not explained carefully enough what analyses we are performing and the results were not reported in detail). The Inparanoid analysis attempted to identify orthologous genes between the *P. micrantha* and *F. vesca* datasets using protein sequences. This analysis revealed 33,127 *P. micrantha* genes that had a putative match in the *F. vesca* gene set (98.6%). However, for the analysis of synteny, we chose only those genes where there was a clear and unambiguous orthologous relationship between the two genomes. This is the reason only 24,555 genes were considered for the synteny analysis. We have amended the text to clarify these points.

Q6. c.) I cannot make much sense out of figure 2B. In the figure resolution I have, individual lines look like they correspond to a single (or very few) gene(s), although I suspect it has to be more genes. It would be good to define the sizes of the blocks somewhere.

R6. We have added the number of genes in each conserved region which are split between pseudomolecules on the *F. vesca* genome. We have also made it clearer that these are only scaffolds that are split between pseudomolecules (i.e. evidence of major rearrangements). There were many instances of individual genes that did not fall into syntenic blocks.

Q7. d.) I'd move figure 9 to the supplementary material and drop sup Figure S1. It still has the same problem as when it was a main figure.

	<p>R7. Done as suggested</p> <p>Reviewer #3: The authors have addressed all my comments and concerns. Q8. I have one minor comment -I found 'CDs' in multiple places. Could you spell it out when it first appears? If it stands for coding DNA sequence, I think 'CDS' is a more common abbreviation.</p> <p>R8. Done as suggested.</p>
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <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>	Yes
<p><b>Resources</b></p> <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>	Yes
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (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.</p>	Yes

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)**

3  
4  
5 3

6  
7 4 **DATA NOTE**

8  
9 5 Matteo Buti<sup>1,7</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>  
10 ([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>  
11 ([luca.natali@unipi.it](mailto:luca.natali@unipi.it)), Matteo Brilli<sup>1,3</sup> ([matteo.brilli.bip@gmail.com](mailto:matteo.brilli.bip@gmail.com)), Alexandre Lomsadze<sup>4</sup>  
12 ([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>  
13 ([lara.giongo@fmach.it](mailto:lara.giongo@fmach.it)), Michael Alonge<sup>5</sup> ([michael.alonge@driscolls.com](mailto:michael.alonge@driscolls.com)), Riccardo Velasco<sup>1</sup>  
14 ([riccardo.velasco@fmach.it](mailto:riccardo.velasco@fmach.it)), Claudio Varotto<sup>1</sup> ([claudio.varotto@fmach.it](mailto:claudio.varotto@fmach.it)), Nada Šurbanovski<sup>1</sup>  
15 ([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>  
16 ([judson.ward@driscolls.com](mailto:judson.ward@driscolls.com)), Kristof Engelen<sup>1</sup> ([engelen.kristof@gmail.com](mailto:engelen.kristof@gmail.com)), Alessandro Cestaro<sup>1</sup>  
17 ([alessandro.cestaro@fmach.it](mailto:alessandro.cestaro@fmach.it)), Andrea Cavallini<sup>2</sup> ([andrea.cavallini@unipi.it](mailto:andrea.cavallini@unipi.it)), Daniel James Sargent  
18  
19 9  
20  
21  
22 10  
23  
24 11  
25  
26  
27 12  
28  
29 13  
30  
31 14  
32  
33  
34 15  
35  
36 16  
37  
38  
39 17  
40  
41 18  
42  
43  
44 19  
45  
46 20  
47  
48  
49 21  
50  
51 22  
52  
53 23  
54  
55  
56 24  
57  
58 25  
59  
60  
61  
62  
63  
64  
65

<sup>1</sup>Fondazione Edmund Mach, Centre for Research and Innovation, via Mach 1, San Michele  
all'Adige, 38010 (TN), 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>4</sup>Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA 30332,  
USA.

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

1 <sup>6</sup>Driscoll's Genetics Limited, East Malling Enterprise Centre, New Road, East Malling, Kent ME19  
2 6BJ, UK.

3  
4 <sup>7</sup>Center for the Development and Improvement of Agri-Food Resources (BIOGEST-SITEIA)  
5  
6  
7 University of Modena and Reggio Emilia, P.le Europa 1, 42124 Reggio nell'Emilia (RE), Italy  
8

9 \*Corresponding Author  
10  
11  
12  
13

## 14 **ABSTRACT**

15  
16  
17 The genus *Potentilla* is closely related to that of *Fragaria*, the economically important strawberry  
18  
19 genus. *Potentilla micrantha* is a species that does not develop berries, but shares numerous  
20  
21 morphological and ecological characteristics with *F. vesca*. These similarities make *P. micrantha* an  
22  
23 attractive choice for comparative genomics studies with *F. vesca*. In this study, the *Potentilla*  
24  
25 *micrantha* genome was sequenced and annotated, and RNA-Seq data from the different  
26  
27 developmental stages of flowering and fruiting were used to develop a set of gene predictions. A 327  
28  
29 Mbp sequence and annotation of the genome of *P. micrantha*, spanning 2,674 sequence contigs, with  
30  
31 an N50 size of 335,712, estimated to cover 80% of the total genome size of the species was developed.  
32  
33  
34 The genus *Potentilla* has a characteristically larger genome size than *Fragaria*, but the recovered  
35  
36 sequence scaffolds were remarkably collinear at the micro-syntenic level with the genome of  
37  
38 *F. vesca*, its closest sequenced relative. A total of 33,602 genes were predicted, and 95.1% of BUSCO  
39  
40 genes were complete within the presented sequence. Thus, we argue that the majority, of the gene-  
41  
42 rich regions of the genome have been sequenced. Comparisons of RNA-Seq data from the stages of  
43  
44 floral and fruit development revealed genes differentially expressed between *P. micrantha* and  
45  
46 *F. vesca*. The data presented are a valuable resource for future studies of berry development in  
47  
48 *Fragaria* and the Rosaceae and they also shed light on the evolution of genome size and organization  
49  
50  
51 in this family.  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

## 7 4 INTRODUCTION

9 5 *Potentilla*, a genus of approximately 500 species [1], is closely-related to that of *Fragaria* [2], the  
10 genera having diverged from a common ancestor just 24 million years ago [3]. The genus *Fragaria*,  
11 a member of the Fragariinae tribe of the Rosaceae family, is economically-important due to the  
12 sweet, aromatic accessory fruits (berries) produced by members of the genus, in particular those of  
13 the cultivated allo-octoploid ( $2n=8\times=56$ ) strawberry species *F. \times ananassa*. The availability of a  
14 genome sequence for a wild diploid relative of the cultivated strawberry, the woodland strawberry  
15 *F. vesca* ( $2n=2\times=14$ ) has enabled the investigation of the molecular basis of many traits of economic  
16 and academic interest in strawberry, including the development of accessory fruits. However, all  
17 members of the *Fragaria* genus produce berries, and as such the use of reverse genetics approaches  
18 to study the genes involved in berry evolution and development would require *Fragaria* mutants that  
19 do not produce fruits, a resource that is not currently available.

20 In the post genomics era comparative analysis permits the study of related, yet divergent species, by  
21 tracing changes at the genomic and transcriptomic levels responsible for their phenotypic differences.

22 Previously, the sequenced genomes of *F. vesca*, *Prunus persica* and *Malus \times domestica* were  
23 compared [4]; the study revealed insights into the evolutionary mechanisms that have shaped the three  
24 species, demonstrating that the *Fragaria* genome underwent significant small-scale structural  
25 rearrangement since it diverged from the common ancestor of the three genera. Comparative  
26 transcriptomics can also be used to reveal differences in the expression of orthologous genes between  
27 organisms at different stages of physiological development [5]. Such an approach suggests that  
28 comparative analyses between *Fragaria* and a closely-related species that does not bear berries may  
29 reveal important insights into the evolution of fruit development. Additionally, species separation is  
30 often related to changes in genome structure, and genome size in particular. Differences in genome



1 size are often the consequence of polyploidization events and/or changes in the abundance of  
2 repetitive DNA, especially transposable elements [6].

3  
4 3 *Potentilla micrantha*, like the majority of species within the *Potentilla* genus does not develop  
5 accessory fruits, but shares numerous morphological characteristics with *F. vesca* (Fig. 1) including  
6  
7 4 plant habit and flower morphology. Notably, they grow within the same ecological niches, and where  
8  
9 5 their ranges of distribution overlap, *P. micrantha* can be found growing nearby populations of  
10  
11  
12 6 *F. vesca* (Sargent, unpublished results). These striking similarities make *P. micrantha* an attractive  
13  
14 7 choice for comparative genomics studies with *F. vesca* to study the genetic basis of berry  
15  
16  
17 8 development in the latter species. As a precursor to a whole genome sequencing initiative, an initial  
18  
19 9 sequencing project focused on the *P. micrantha* chloroplast was undertaken using the Illumina HiSeq  
20  
21  
22 10 and PacBio RS sequencing platforms [7].  
23  
24 11  
25  
26  
27 12

## 28 **DATA DESCRIPTION**

29 13  
30  
31 14 The objectives of this study were to develop a genomic toolkit for *P. micrantha* to permit comparative  
32  
33  
34 15 genomic and transcriptomic studies with *F. vesca*, with a view to identifying the evolutionary changes  
35  
36 16 that have occurred between the two species. The genome size of *P. micrantha* was determined by  
37  
38  
39 17 flow cytology and the nuclear genome was sequenced and assembled from Illumina and PacBio  
40  
41 18 sequencing reads, assembled and integrated using ALLPATHS and PBJelly. Gene predictions from  
42  
43  
44 19 the *P. micrantha* genome were made with support of RNA-Seq data generated from tissue libraries  
45  
46 20 sampled during flower and fruit development. The genome of *F. vesca* was compared to the  
47  
48  
49 21 sequencing scaffolds produced for *P. micrantha*, and whilst they exhibited a remarkable degree of  
50  
51 22 collinearity at the micro-syntenic level, large-scale differences in transposon activity were identified  
52  
53  
54 23 that could be responsible for the large differences in genome size between the two species. The dataset  
55  
56 24 we report will be useful for comparative studies of a number of traits between *P. micrantha* and its  
57  
58 25 economically-important close relatives.  
59  
60  
61 26

# 1 RESULTS

## 2 2 Flow cytometry, heterozygosity estimation and genome assembly

3  
4 3 DNA was extracted from *Potentilla micrantha* young, unexpanded leaves. Flow cytometry using a  
5  
6  
7 4 *V. minor* internal standard with a DNA content of 1.52 pg/2C returned average DNA quantities of  
8  
9  
10 5 0.52 pg/2C for *F. vesca* ‘Hawaii 4’ and 0.83 pg/2C for *P. micrantha* over three biological replicates.  
11  
12 6 Using the calculation of Dolezel et al. (2003) [8] that 1 pg DNA is equivalent to 978 Mbp of DNA  
13  
14 7 sequence, the genome size of *P. micrantha* was determined as 405.87 Mbp in length whilst that of  
15  
16  
17 8 *F. vesca* ‘Hawaii 4’ was calculated to be 254.28 Mbp.

18  
19 9 Data were returned for the overlapping fragment library (OLF) and all four mate-pair libraries  
20  
21  
22 10 sequenced using Illumina HiSeq. In total, 61.4 Gbp of data were returned and the relative depth of  
23  
24 11 coverage obtained for the *P. micrantha* genome from each library is given in Additional File 1: Table  
25  
26  
27 12 S1. Four different PacBio RS sequencing libraries were constructed and sequenced using two  
28  
29 13 different versions of the PacBio chemistry (Additional File 2: Table S2). From the sequencing of 63  
30  
31  
32 14 SMRT cells, 6,447,413 sequences with an average length of 2,221 bp were recovered, totaling  
33  
34 15 14.32 Gb of long read sequence data. From the data, 33× equivalent of sequence was contained in  
35  
36 16 reads longer than 1 kb which were used for gap filling of the Illumina assembly using PBJelly [9].  
37

38  
39 17 The initial ALLPATHS assembly of the Illumina short-read sequences produced 33,026 contigs with  
40  
41 18 an N50 of 16,235 bp and a total length of 247,565,733 bp. Following scaffolding, a genome assembly  
42  
43  
44 19 with a total length of 315,266,043 bp contained in 2,866 sequencing scaffolds was returned. The final  
45  
46 20 scaffold set returned following ALLPATHS assembly contained a total of 0.07% ambiguous sites  
47  
48  
49 21 (SNPs), revealing the genome of *P. micrantha* to be one of the most homozygous naturally-occurring  
50  
51 22 genomes sequenced to date. Following incorporation of the PacBio RS data using PBJelly [9], the  
52  
53  
54 23 *P. micrantha* sequence assembly contained 326,533,584 bp of sequence data, a 3.5% increase over  
55  
56 24 the ALLPATHS Illumina assembly, in 2,674 scaffolds. The longest and N50 scaffold lengths both  
57  
58 25 increased following gap filling by 9.3% and 5.1% respectively, but most significantly, the number of  
59  
60  
61 26 gapped Ns in the assembly was reduced by 59.7% to 27,311,787 (8.4% of the final assembly) (Table

1) The final scaffolded assembly contained 80.45% of the total estimated genome size for *P. micrantha* as calculated by flow cytometry. Sequence scaffold size ranged from 935 bp to 3,488,351 bp. Of the 2,674 scaffolds, 878 (32.8%) were less than 10 kbp in length, 534 (20%) were between 10 and 50 kbp in length, 738 (27.6%) were between 50 and 200 kbp in length, 500 (18.7%) contained between 200 kbp and 1 Mbp of sequence, and the remaining 23 (0.9%) contained over 1 Mbp of sequence. The majority of the 1,440 benchmarking single-copy orthologous (BUSCO) groups queried [10] were present in the genome sequence, with 95.1% (1,337 complete and single copy and 33 complete and duplicated BUSCOs) identified within the sequencing scaffolds.

## Gene prediction and preliminary annotation

The results of the combined alignment of the 12 RNA-seq read sets to the *Potentilla* genome sequence scaffolds and number of splice sites identified using STAR is presented in Additional File 3: Table S3. A total of 1,908 consensus repeat sequences were generated by RepeatModeler totaling 1,431,262 bp and having a GC content of 40.8%. The total ATCG content of sequencing scaffolds greater than 10 kb in length was 298,987,576 bp. A total of 138,597,969 bp (46.36%) of the genome sequence were masked using the consensus sequences in the RepeatModeler library, including 26,359 (7.5%) of the mapped GT-AG introns identified by STAR. Gene prediction using GeneMark-ET on the masked genome identified a total of 33,602 genes, of which 32,137 were predictions containing multiple exons, and 4,655 were single exon predictions. A total of 172,791 exons were predicted, with an average length of 223 bp and an average of 5.14 exons per gene. A total of 139,216 introns were predicted in the CDS of the genes, with an average intron length of 499 bp. BUSCO analyses were compared between the gene predictions developed for *P. micrantha* and those of *F. vesca*. In total, 1,282 (89%) complete and 68 (4.7%) fragmented BUSCOs (93.75% total) were recovered for *P. micrantha*, compared to 1,303 (90.5%) complete and 79 (5.5%) fragmented BUSCOs (95.6%) recovered for *F. vesca* gene predictions indicating a similar level of completeness of the *P. micrantha*

1 to its nearest sequenced relative. Following a local BLAST search and BLAST2GO analysis, a total  
2 of 27,968 *P. micrantha* predicted genes were assigned a preliminary gene annotation.  
3

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

5 Following the inparanoid analysis, a total of 33,127 genes returned an orthologous relationship with  
6 one or more *F. vesca* gene predictions at the amino acid level (98.6%). A subsequent BLAST analysis  
7 of the gene predictions against the *F. vesca* v2.0 pseudomolecules identified a total of 24,641  
8 *P. micrantha* genes that returned an unambiguous match with the position with an orthologous gene  
9 on the *F. vesca* genome. A total of 1,682 *P. micrantha* sequence scaffolds, containing 315,081,089  
10 bp (96.5% of the total sequence) contained at least one gene that was anchored to one of the *F. vesca*  
11 v2.0 pseudomolecules. Of those, 573 contained at least ten orthologous gene sequences, 118 contained  
12 at least 50 orthologous sequences and 32 contained over 100 orthologous (Supplementary Excel File  
13 1). Scaffold 'Contig145', the largest scaffold in the *P. micrantha* genome sequence (3,488,351 bp)  
14 contained the largest number of orthologous gene sequences anchored to the *F. vesca* v2.0 genome  
15 sequence (560), whilst scaffold 'Contig2191' was the smallest anchored scaffold at 1,163 bp, and  
16 containing a single orthologous gene sequence. Comparison of the two genomes revealed a  
17 remarkable degree of micro-synteny with majority of the *P. micrantha* scaffolds spanning  
18 uninterrupted regions of the *F. vesca* genome sequence (Data not shown). A very high degree of  
19 collinearity in gene order was observed between *P. micrantha* scaffolds and the *F. vesca*  
20 pseudomolecules (Fig. 2a). In general, only a small number of inversions were observed between  
21 syntenic blocks between the two genomes, and just eight *Potentilla* scaffolds contained distinct  
22 syntenic blocks that aligned with more than one *Fragaria* pseudomolecule (Fig. 2b). Scaffold  
23 anchoring to a genetic map however was not performed for the *P. micrantha* genome sequence, and  
24 as such, a comparison of macrosynteny between *Fragaria* and *Potentilla* could not be made.  
25

#### 26 **Gene expression during fruit development**

1 Tissues from five stages of flowering and ‘fruit’ development were harvested from *Potentilla*  
2 *micrantha* untreated flowers in biological duplicates or triplicates for RNA isolation. The stages of  
3  
4  
5 3 flowering followed those identified in *Fragaria* by [11], with the addition of a stage 0 (unopened  
6  
7 4 flowers) and young unexpanded leaf tissue. The selected developmental stages are shown in Fig. 3.  
8  
9  
10 5 RNA-libraries were made and sequenced with Illumina HiSeq2000. Following QC and adapters  
11  
12 6 trimming, a total of 619,085,115 101 bp paired reads were obtained from the 12 *P. micrantha* RNA-  
13  
14 7 seq libraries. Sequencing yield from individual libraries ranged from 29,653,058 to 60,158,302 reads  
15  
16  
17 8 per sample (Additional File 4: Table S4). Following trimming, the number of reads available for  
18  
19 9 *Fragaria* from the published sequences of [11] were 1,236,882,540, with reads per library ranging  
20  
21  
22 10 from 109,643,225 to 155,643,061. Between 62% and 69% of *P. micrantha* filtered reads per library  
23  
24 11 mapped to the *P. micrantha* gene prediction set, whilst similarly 63% to 67% of *F. vesca* filtered  
25  
26  
27 12 reads per library mapped to the *F. vesca* gene prediction set (Additional File 4: Table S4). A total of  
28  
29 13 1,556 genes were differentially expressed between the four developmental stages in at least one pair-  
30  
31  
32 14 wise comparison of the different stages in *P. micrantha*, whilst 816 genes were differentially  
33  
34 15 expressed in *F. vesca* between the four stages (Fig. 4). A total of 52.44% and 43.38% of the  
35  
36 16 differentially expressed genes were GO-annotated for *P. micrantha* and *F. vesca* respectively  
37  
38  
39 17 (Additional File 5: Fig. S1). Analysis of the GO terms for *F. vesca* and *P. micrantha* revealed an  
40  
41 18 enrichment for genes associated with lipid metabolic processes, transporter activity, and transcription  
42  
43  
44 19 factor activity and transcription regulator activity in *F. vesca* over *P. micrantha* (Fig. 5). The gene  
45  
46 20 expression profiles between the four developmental stages studied in the two species showed no clear  
47  
48  
49 21 consistent patterns between the two species overall (Additional File 6: Fig. S2), however the common  
50  
51 22 differentially expressed genes displayed largely similar expression patterns (Fig. 6), with some  
52  
53  
54 23 exceptions, most notably gene1369-v1.0-hybrid and its homologue in *P. micrantha* (17717\_t), a  
55  
56 24 predicted 3-hydroxy-3-methylglutaryl coenzyme A reductase 1, which was highly expressed in  
57  
58 25 *F. vesca* but exhibited far lower levels of gene expression in *P. micrantha*.

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

1  
2 2 A total of 75 *P. micrantha* and 81 *F. vesca* predicted proteins containing MADS-box conserved  
3  
4 3 domains were aligned and phylogenetic trees were constructed to reliably identify orthology  
5  
6 4 relationships between *P. micrantha* and *F. vesca* genes. The three methods employed for phylogenetic  
7  
8 5 reconstruction (ML, MP, NJ) returned largely congruent topologies for the nodes with more than 50%  
9  
10 6 bootstrap support, with NJ providing a slightly more resolved tree given the use of a pairwise, instead  
11  
12 7 of a partial deletion approach. Fig. 7 displays the ML phylogenetic reconstruction of the *P. micrantha*  
13  
14 8 and *F. vesca* genes containing MADs-box, along with the gene expression levels for each gene (data  
15  
16 9 for the NJ and MP trees are not shown). The majority of the genes were retained after the divergence  
17  
18 10 of the species, indicated by a large proportion of orthologous pairs retrieved. Only a few events of  
19  
20 11 lineage-specific gene loss/duplication were observed. Both observations are in line with the lack of  
21  
22 12 ploidy changes within *P. micrantha* and *F. vesca* in the estimated 24.22 million years since species  
23  
24 13 divergence. As expected, the majority of orthologous pairs shared similar expression patterns. Based  
25  
26 14 on the ML gene tree however, three clades of orthologous genes were identified that were not  
27  
28 15 expressed, or poorly expressed in *P. micrantha* but highly expressed in *F. vesca* (Fig. 8). The three  
29  
30 16 clades, numbered as 1, 2 and 3 on Fig. 8, contained the following genes: clade 1 contained genes  
31  
32 17 27280\_t (*P. micrantha*) and gene25871-v1.0-hybrid (*F. vesca*), which displayed highest homology to  
33  
34 18 *A. thaliana* AGL36, a sequence-specific DNA binding transcription factor active during endosperm  
35  
36 19 development [12]; clade 2 contained genes 26598\_t (*P. micrantha*) and gene18483-v1.0-hybrid  
37  
38 20 (*F. vesca*), whose closest *A. thaliana* homologue was AGL62, a MADS gene that promotes embryo  
39  
40 21 development, indicating an essential role of endosperm cellularization for viable seed formation [13];  
41  
42 22 and clade 3 contained *P. micrantha* genes 23638\_t, 23641t and 759\_t and *F. vesca* genes gene32155-  
43  
44 23 v1.0-hybrid and gene13277-v1.0-hybrid, whose closest *A. thaliana* homologue AGL15 delays  
45  
46 24 senescence programs in perianth organs and developing fruits and alters the process of seed  
47  
48 25 desiccation [14].  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

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

1  
2 2 In total, 1,001,838 of 1,484,780 reads clustered with RepeatExplorer were grouped into 107,190  
3  
4 3 clusters, representing 67.5% of the genome. No predominant repeat families were identified in the  
5  
6  
7 4 *P. micrantha* genome, with the most redundant repeat cluster representing just 1.18% of the total  
8  
9  
10 5 genome length. LTR-retrotransposons made up the main fraction (24.1%) of the *P. micrantha*  
11  
12 6 genome (Additional File 7: Fig. S3), with a *Gypsy* to *Copia* ratio of approximately 2:1. Terminal-  
13  
14 7 repeat retrotransposons in miniature (TRIMs) were poorly represented, making up just 0.2% of the  
15  
16  
17 8 genome, whilst putative DNA transposons accounted for 5.7% of the genome and included putative  
18  
19 9 CACTA, Harbinger, and hAT elements, with other, unclassified repeats accounting for 10.6% of the  
20  
21  
22 10 genome. A comparison of the repetitive portion of the *F. vesca* and *P. micrantha* genomes performed  
23  
24 11 by pairwise clustering of Illumina sequence reads revealed significant diversification between the  
25  
26  
27 12 repetitive component of the genomes of the two species (Additional File 8: Fig. S4). Among the top  
28  
29 13 291 repeat clusters that had a genome proportion >0.01%, 107 were specific to *P. micrantha*, 51 were  
30  
31  
32 14 specific to *F. vesca*, whilst only 25 were similarly represented in the two species. Among all repeat  
33  
34 15 classes, only ribosomal DNAs show similar genome proportions between *P. micrantha* and *F. vesca*.

35  
36 16  
37

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

39 17  
40  
41 18 Of the 505 LTR-REs characterised, 220 (43.6%) belonged to the *Copia* superfamily, with the greatest  
42  
43  
44 19 proportion belonging to the *Bianca* family, 256 (50.7%) belonged to the *Gypsy* superfamily, with the  
45  
46 20 greatest proportion belonging to the *Ogre/TAT* family, whilst the remaining 29 (5.7%) could not be  
47  
48  
49 21 placed into a specific superfamily. Table 2 lists the proportion of the annotated 505 LTR-REs in each  
50  
51 22 superfamily, and the numbers of elements contained in each sub-family within the *Copia* and *Gypsy*  
52  
53  
54 23 super-families. For RE insertion age determination, the mean synonymous substitution rate between  
55  
56 24 *P. micrantha* and *F. vesca*, was estimated by comparing 50 orthologous genes, which equated to  
57  
58 25 52,703 bp of aligned sequences between the two species, resulting to be 0.064 synonymous  
59  
60  
61 26 substitutions per site ( $K_s$ ). Using a timescale of 24.22 million years since the separation of

1 *P. micrantha* and *F. vesca*, and a  $K_s$  of 0.064, the resulting synonymous substitution rate was  
2 2.64×10<sup>-9</sup> substitutions per year. As mutation rates for LTR retrotransposons have been estimated to  
3  
4 be approximately two-fold higher than silent site mutation rates for protein coding genes (SanMiguel  
5 and Bennetzen 1998; Ma and Bennetzen 2004), a substitution rate per year of 5.28×10<sup>-9</sup> was used in  
6  
7 calculations of LTR-RE insertion dates. When the whole set of usable retrotransposons was taken  
8  
9 into account, the nucleotide distance (K) between sister LTRs showed a large degree of variation  
10  
11 between retro-elements, ranging from 0 to 0.124 using the Kimura two parameter method, which  
12  
13 represents a time span of at most 23.54 million years.  
14  
15  
16  
17  
18  
19  
20  
21

## 22 **DISCUSSION**

23  
24 In this investigation, we present a set of resources for *P. micrantha*, which will form the foundation  
25  
26 for future genomics studies in the species. Here, the genome of *P. micrantha*, a member of the  
27  
28 Rosaceae, a diverse family of fruiting perennial plant genera, was sequenced using both short-read  
29  
30 Illumina and long-read PacBio sequence data, and the resulting data was assembled into a highly  
31  
32 contiguous reference sequence for the genus *Potentilla*. The study has provided a foundational  
33  
34 resource for the future comparative genomics studies within the Rosoideae sub-family of Rosaceae  
35  
36 in particular, but also in the Rosaceae as a whole. PacBio data (using early iterations of the sequencing  
37  
38 chemistry) were proficiently integrated with short-reads, significantly improving the contiguity of the  
39  
40 assembly; however the PacBio throughput was not sufficient to permit independent *de novo* assembly.  
41  
42 Nonetheless, whilst fragmented, the genome and sequence presented here have a quality similar to  
43  
44 the *F. vesca* genome, containing significantly fewer un-sequenced gaps within scaffolds, and is far  
45  
46 more contiguous than that of *R. occidentalis* [15]. Along with the set of gene predictions presented,  
47  
48 it represents a valuable resource for studying the genetic basis of a number of key morphological  
49  
50 traits that differ between *P. micrantha* and its closest sequenced relatives.  
51  
52  
53  
54  
55  
56

57  
58 *Potentilla* and *Fragaria* are separated by just 24.22 million years of evolution [3], however, in this  
59  
60 investigation, we show the genome of *P. micrantha* is 59.6% larger than that of *F. vesca*, and it is  
61  
62  
63  
64  
65



1 also larger than the available genomes of the other Fragariinae i.e. *Rubus* [16,17] and *Rosa* species  
2 [18,19] to which it is more distantly related. We also demonstrate here that *P. micrantha* and *F. vesca*  
3 exhibit a remarkable degree of microsynteny of the coding portion of the genome, with the main  
4 differences being short-range inversions. Nonetheless, the apparent differences in insertion age of  
5 transposable elements in the two genomes has led to significant differences in the repetitive portions.  
6  
7 Whereas the genome structure of *P. micrantha* is similar to that of most angiosperm species [20],  
8 with a repetitive component amounting to around 41.5% of the total genome content, the genome of  
9 *F. vesca* has been previously demonstrated to contain just 22% repetitive elements [21]. Contrary to  
10 the coding or non-repetitive genome, the repetitive fractions of the *P. micrantha* and *F. vesca*  
11 genomes are highly diversified, suggesting that the overwhelming majority of retrotransposon activity  
12 in the genus *Potentilla* occurred after the divergence of the two genera from their common progenitor.  
13  
14 The data presented here strongly indicate that retrotransposon activity (or the lack thereof in the genus  
15 *Fragaria*) is responsible for the significant difference between the genome size of *Fragaria* and its  
16 closest relatives, and support the assertion of Potter et al. (2007) [2] that *Fragaria* should be treated  
17 as a distinct genus, separate from *Potentilla*.  
18  
19 Gene expression patterns for differentially expressed genes that were common to both *F. vesca* and  
20 *P. micrantha* were largely similar between the two species, however one gene, a 3-hydroxy-3-  
21 methylglutaryl coenzyme A reductase 1 homologue displayed significantly higher gene expression  
22 levels in *F. vesca*. The 3-hydroxy-3-methylglutaryl coenzyme A reductase 1 gene catalyzes the first  
23 committed step in the cytosolic isoprenoid biosynthesis pathway [22]. Loss of function mutants of  
24 this gene in *Arabidopsis* display a dwarf phenotype due to suppression of cell elongation and reduced  
25 sterol levels [22]. Sterols are precursors in cellulose synthesis, important for cell-wall formation [23]  
26 and fruit development, and as such, up-regulation in the 3-hydroxy-3-methylglutaryl coenzyme A  
27 reductase 1 gene during fruit development in *F. vesca* over *P. micrantha* may indicate a role for this  
28 enzyme in berry formation in *Fragaria*.

1 In contrast to the gene expression patterns of differentially expressed genes common to both *F. vesca*  
2 and *P. micrantha* during fruit development, global patterns of gene expression during fruit  
3 development differed between the two species. The gene ontology for the *F. vesca* expression profile  
4 was enriched for genes with transcription factor and transcription regulator activity as well as  
5 transporter activity and lipid metabolic processes. A study of the differences in transcriptional  
6 regulation between *F. vesca* and *P. micrantha* therefore may provide clues to the genetic basis of  
7 berry formation in *F. vesca*. MADS-box transcription factors have been implicated in a wide and  
8 extremely diverse array of developmental processes in plants [24], and were initially demonstrated to  
9 play a major role in floral organ differentiation, including gametophyte, embryo and seed  
10 development, as well as flower and fruit development. A study of the differential expression of  
11 MADS-box genes revealed three clades of orthologous genes where gene expression of orthologous  
12 genes was up-regulated in *F. vesca* with respect to *P. micrantha*. One clade contained genes that were  
13 homologous to AGL36, a transcription factor crucial for endosperm differentiation and development  
14 [12,25]. Another clade contained genes homologous to *A. thaliana* AGL62, which likewise has been  
15 implicated in embryo development, and is thought to have an essential role of endosperm  
16 cellularization for viable seed formation [13]. The third clade contained genes homologous to AGL15  
17 reported to have diverse roles in embryogenesis, fruit maturation, seed desiccation and the repression  
18 of floral transition [14,26], as well as being a positive regulator of the expression of mir156, a  
19 repressor of floral transition [27].  
20  
21 The set of genomics tools developed here for a non-fruiting relative of *F. vesca*, including a genome  
22 sequence, gene predictions and RNA-Seq data is a valuable foundational resource for more detailed  
23 future functional studies of fleshy receptacle or berry development.

## 24 **METHODS**

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

1 A specimen of *Potentilla micrantha* was collected from Avala, Serbia in spring 2012 and  
2 subsequently used for sequencing. The plant was maintained in a growth room at a constant  
3 temperature of 24 degrees during the day and 18 degrees at night, with a 16-hour photoperiod to  
4 encourage new shoot development. Young leaves were harvested and subjected to flow cytometry by  
5 Plant Cytometry Services, NL. Measurements were taken in triplicate against a *Vicia minor* internal  
6 standard using the propidium iodide fluorescent dye. The *F. vesca* accession ‘Hawaii 4’ for which a  
7 whole genome sequence has been published [21] was analyzed for comparison. Prior to harvesting  
8 leaf material for DNA extraction, the plant was moved to a darkened growth chamber for 120 hours,  
9 maintaining a constant temperature of 22 degrees. DNA was extracted from young, unexpanded leaf  
10 material using the modified CTAB extraction protocol of Chen and Ronald (1999) [28], quantified  
11 using a Nanodrop spectrophotometer and Qubit fluorometer, and assessed for integrity by agarose gel  
12 electrophoresis against a  $\lambda$  *HindIII* size standard.

13 Since *P. micrantha* does not reproduce asexually from runners, a seedling population obtained from  
14 the selfing of the original mother plant was maintained from which to harvest tissue from stages of  
15 floral and fruiting development. Flowers of *P. micrantha* and *F. vesca*, along with two other  
16 *Potentilla* species, *P. reptans* and *P. indica* were treated with naphthaleneacetic acid (NAA; Sigma-  
17 Aldrich), N-1-naphthylphthalamic acid (NPA; Sigma-Aldrich), gibberellic acid (GA3; Sigma-  
18 Aldrich) and a combination of NAA and NPA, following the methods of Kang et al. (2013) [11].  
19 Briefly, stock solutions of 50 mM NAA, 50mM NPA, and 100mM GA3 were made in ethanol and  
20 diluted with two drops of Tween 20 and water before application. The final treatment concentrations  
21 were 500  $\mu$ M for NAA and GA3 and 100  $\mu$ M for NPA. 50 ml of hormone solution was pipetted onto  
22 the receptacle of each emasculated flower every two days for twelve days.

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

24 Tissues from five stages of flowering and ‘fruit’ development were harvested from untreated flowers  
25 in biological duplicates or triplicates for RNA isolation. The stages of flowering followed those

1 identified in *Fragaria* by Kang et al. (2013) [11], with the addition of a stage 0 (unopened flowers)  
2 and young unexpanded leaf tissue. The selected developmental stages are shown in Fig. 3. RNA was  
3 extracted from 50 mg of snap-frozen tissue from each developmental stage using the Spectrum plant  
4 total RNA extraction kit (Sigma) with an on-column DNase I digestion (Sigma) step. The extraction  
5 protocol followed the manufacturers' recommendations with two minor modifications: 1% PVP was  
6 added to the lysis solution, and the number of washes at each stage was doubled (i.e. two washes were  
7 performed with wash solution 1 and four washes were performed with wash solution 2). The RNA  
8 extracted from each sample was diluted in 50 µl of elution solution (Sigma). Following elution, total  
9 RNA was quantified using a Nanodrop spectrophotometer and Qubit fluorometer and assessed for  
10 integrity using a Bioanalyzer (Agilent). Samples returning a RIN value greater than 7.5 were  
11 considered acceptable for sequencing. A total of 12 Illumina TruSeq libraries were constructed from  
12 2 µg of total RNA. Libraries were made from the following samples; one from stage 0, two from  
13 stage 1, two from stage 2, three from stage 3 and three from stage 4. A final library was made from  
14 RNA of young leaf tissue. The libraries were sequenced in triplex per single lane of Illumina  
15 HiSeq2000. Samples were indexed and multiplexed, and then 101 bp paired-end sequencing was  
16 performed using the Illumina HiSeq 2000 platform at the Weill Medical core genomics facility of  
17 Cornell University.

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

19 A strategy following the ALLPATHs-LG protocol was followed to produce an initial assembly using  
20 second-generation sequence data. Five sequencing libraries were developed; an overlapping fragment  
21 library (OLF) with an insert size of 170 bp, and four libraries of 3 kb, 5 kb, 8 kb and 12 kb. The OLF  
22 library was created using the Illumina Nextera library preparation kit following the manufacturers'  
23 recommendations and was sequenced in simplex on a single lane of Illumina HiSeq2000, whilst the  
24 MP libraries were prepared using the Illumina Mate Pair Library v2 kit following the manufacturers'  
25 recommendations and were subsequently sequenced in duplex. All sequencing was performed at the

1 Weill Medical Centre core genomics facility at Cornell University. ALLPATHS-LG [29] was run  
2 using the sequencing libraries described above using default settings. Subsequently, a selection of  
3 SMRT-bell sequencing libraries were constructed using various versions of the PacBio RS  
4 sequencing kits and chemistries (Additional File 2: Table S2) and PBJelly [9] running default settings  
5 was used to incorporate data generated using the PacBio RS platform (Pacific Biosciences) into the  
6 ALLPATHS-LG Illumina assembly scaffolds. Identification of benchmarking universal single-copy  
7 orthologs (BUSCOs) was performed using BUSCO v3 [10] running default parameters and using  
8 1,440 BUSCO groups from the embryophyta\_odb9 (plant) lineage data.

## 9 **Gene prediction, annotation, determination of gene orthology and evaluation of synteny** 10 **between *Potentilla* and *Fragaria* genomes**

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

24 Orthologous relationships between *Fragaria* and *Potentilla* genes was determined through sequence  
25 clustering performed using Inparanoid 7 [35]. Analyses were based only on homology, as an

1 alternative to the more stringent ortholog classification. *Prunus persica* v2.0.a1 predicted proteins  
2 downloaded from the GDR [36] and *Potentilla micrantha* and *Fragaria vesca* protein sequences were  
3  
4  
5 3 blasted all against all and the output file was filtered at the following thresholds: maximum E-  
6  
7 4 value= $10^{-4}$  and query coverage of at least 50%. The resulting file was used as an input to the MCL  
8  
9  
10 5 algorithm using as edge weight  $-\log_{10}(\text{evaluate})$  (all E-values=0 were changed to 1E-300). To explore  
11  
12 6 more thoroughly the homology network used as input, the MCL algorithm was run at different  
13  
14 7 granularity levels (inflation parameter equal to 1.5, 1.7, 2.0, 2.3, 2.4, 2.7, 3) and then a table indicating  
15  
16  
17 8 cluster memberships at the different stringencies was compiled for each node. Ortholog classification  
18  
19 9 was produced using Inparanoid 7 [35] for pairs of species in all combinations. The resulting sqltables  
20  
21  
22 10 were then used as an input for QuickParanoid (<http://pl.postech.ac.kr/QuickParanoid/>) and the  
23  
24 11 sequences were combined in a three-species ortholog classification. The clusters obtained with  
25  
26  
27 12 QuickParanoid were used to calculate the number of genes contained in each cluster for both  
28  
29 13 *Potentilla* and *Fragaria*.  
30  
31 14 *Potentilla* gene predictions for which an orthologous relationship was identified through the  
32  
33  
34 15 inparanoid analysis, were used as queries to identify the physical locations of orthologous sequences  
35  
36 16 on the *F. vesca* v2.0 pseudomolecules and those sequences that returned a single, unambiguous match  
37  
38  
39 17 on the genome sequence were used to evaluate synteny between the two species. Since the *Potentilla*  
40  
41 18 genomic scaffolds were not oriented and ordered against a reference genetic map, conservation of  
42  
43  
44 19 synteny between the *Potentilla* and *Fragaria* genomes was determined through a comparison of the  
45  
46 20 physical positions of orthologous gene sequences on the sequence scaffolds of *Potentilla* and the  
47  
48  
49 21 pseudomolecules of *Fragaria*. Criteria for the identification of syntenic regions followed that of Jung  
50  
51 22 et al (2012). No attempt was therefore made to infer macro-syntenic structure on a chromosome scale  
52  
53 23 between the two genomes.

54  
55  
56 24  
57  
58 25 **Gene expression during stages of fruit development in *Potentilla micrantha* and *Fragaria vesca***  
59  
60  
61  
62  
63  
64  
65

1 The quality of the raw reads generated as described above was checked with FastQC [37];  
2 Trimmomatic [38] was used to remove adapter sequences. The *F. vesca* .sra files [11] were used to  
3  
4 compare gene expression in *Fragaria* with *Potentilla*; *Fragaria* reads from the same developmental  
5  
6 stage were merged and treated as a single data set since data from *Potentilla* was not generated from  
7  
8 individual floral organs. The 12 trimmed *P. micrantha* RNA-seq libraries were mapped on the  
9  
10 *P. micrantha* gene prediction CDS, while the ten *F. vesca* sets were mapped to the *F. vesca* v1.0 gene  
11  
12 prediction CDS [21] downloaded from the GDR [36] using Bowtie2 [39] and default settings. The  
13  
14 number of reads mapping to each gene for each RNA set was calculated from the .sam alignment files  
15  
16 derived from Bowtie2.  
17  
18  
19 Counts of RNA-seq reads over transcripts were used to calculate the gene expression level in  
20  
21  
22  $FPKM=10^9*ER/(EL \times MR)$ , where ER was the number of mapped reads in the exons of a particular  
23  
24 gene, EL was the sum of exon length in base pairs, and MR was the total number of mapped reads  
25  
26 [40]. FPKM was used to distinguish expressed genes from inactive genes (those not returning any  
27  
28 expression data) during the flower development in each species. Further, FPKM was used to define  
29  
30 a set of highly expressed genes: Genes were considered as ‘highly-expressed’ if  $FPKM>1000$ . Genes  
31  
32 that returned an  $FPKM<1000$  in all samples were removed from further differential expression  
33  
34 analysis. The retained differentially expressed genes were processed by performing a linear rescaling  
35  
36 of the  $\log_2$ -counts, aligning the distributions for every sample at their distribution modes, followed  
37  
38 by variance stabilization to ensure homoscedasticity. A one-way ANOVA was performed gene-by-  
39  
40 gene on the rescaled  $\log_2$ -counts to detect changes in expression among different developmental  
41  
42 phases. Differentially expressed genes (DEGs) were selected by setting cutoffs both on the p-values  
43  
44 from the ANOVA F-tests, as well as on the magnitude of observed changes represented by the square  
45  
46 root of the ANOVA MSR values (equivalent to using volcano plots for two-condition studies). Genes  
47  
48 were considered differentially expressed if the  $\sqrt{MSR} > 2.00$  and p-value  $< 10^{-3}$ .  
49  
50  
51 Gene Ontology enrichment analysis of DEGs sets of *Potentilla micrantha* and *Fragaria vesca* was  
52  
53 carried out using Blast2GO 2.8.0 [41] with “Fisher’s exact test” method, considering as “enriched”  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 the GO categories with  $FDR < 0.05$ . *Potentilla micrantha* whole transcriptome functional annotation  
2 obtained in this work was used as background for *Potentilla* GO enrichment analysis, while the  
3  
4  
5 “InterPro GO for GeneMark hybrid transcripts” database downloaded from GDR website was used  
6  
7 as background for *Fragaria vesca*. Cytoscape 3.5.1 [42] with the BiNGO 3.0.3 plugin was used for  
8  
9 the GO-slim network visualization of enriched GO categories over *Fragaria vesca* and *Potentilla*  
10  
11  
12 *micrantha* DEGs. For determination of over-representation, the Benjamini and Hochberg FDR-  
13  
14 adjusted significance level cutoff was 0.05.  
15  
16  
17  
18  
19

## 20 **Phylogenetic and functional analysis of MADs-box domain-containing genes and gene** 21 22 **expression profile mapping**

23  
24 Protein sequences of *Potentilla* (this publication) and *Fragaria* (Fvesca\_v1.0\_hybrid;  
25  
26  
27 [www.rosaceae.org](http://www.rosaceae.org)) were analysed on the NCBI conserved domain database [43]. All proteins  
28  
29 containing a MADS-box domain were retrieved and the MADS-box extracted with Bedtools getfasta  
30  
31  
32 [44] using default parameters. An initial sequence alignment was carried out using ClustalW and  
33  
34 pairwise distances were calculated to eliminate outliers. A total of 16 sequences were removed from  
35  
36  
37 further analysis since they were too short and possessed incomplete N-terminal ends, indicating they  
38  
39 were likely pseudogenes. The alignment used for phylogenetic analysis was constructed with SATé-  
40  
41  
42 II [45] and contained 156 protein sequences (75 from *Potentilla* and 81 from *Fragaria*).

43  
44 Three methods, Maximum Likelihood (ML), Maximum Parsimony (MP) and Neighbour-joining  
45  
46 (NJ), each with 1,000 bootstrap replicates were employed for phylogenetic reconstruction of the  
47  
48  
49 MADS-box domain containing genes using Mega 7.0.14 [46]. Where missing data was present in the  
50  
51 alignment, deletion of columns containing a fraction of missing data above 10% and 30% was  
52  
53  
54 performed for ML and MP methods. Pairwise deletion was instead used in the case of NJ, to maximise  
55  
56 the phylogenetic information retained in the alignment. The ML topology was used as reference for  
57  
58  
59 further analysis.  
60  
61  
62  
63  
64  
65



1 The expression profiles of the genes containing a MADS-box were used to decorate the phylogenetic  
2 tree using iTOL v2 [47], allowing the identification of orthologous MADS-box gene pairs displaying  
3 differential gene expression profiles between *Potentilla* and *Fragaria*. Curated annotation of  
4  
5 differentially expressed putative gene function was carried out using BLASTp homology searches of  
6  
7 the TAIR database [48].  
8  
9  
10

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

16 To identify and characterize genomic repeats in the *P. micrantha* genome, a reduced set of 2,000,000  
17  
18 randomly selected genomic Illumina reads, corresponding to 0.57× of the *P. micrantha* genome were  
19  
20 subjected to clustering using RepeatExplorer [49]. Among the clusters produced, the top clusters,  
21  
22 with a genome proportion higher than 0.01%, were annotated using 0.2 as cutoff for cluster  
23  
24 connection through mates. Clusters that were annotated as similar to phi-X174 were removed as  
25  
26 contaminants. The output of RepeatExplorer was also used to prepare an in-house library containing  
27  
28 all contigs belonging to clusters annotated by RepeatExplorer as long terminal repeat retrotransposons  
29  
30 (LTR-REs) by similarity search against RepBase [50]. Subsequently, pairwise hybrid clustering  
31  
32 between a random set of 1,431,114 Illumina reads derived from *P. micrantha* genomic DNA and  
33  
34 1,090,102 *F. vesca* genomic reads, each corresponding to 0.41× of the respective genomes was  
35  
36 performed using RepeatExplorer [49].  
37  
38  
39  
40  
41  
42  
43  
44  
45

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

47 LTR-FINDER [51] was used to isolate putative full-length LTR-REs from 280 randomly-selected  
48  
49 *Potentilla* genome sequence scaffolds and alignment boundaries were obtained by adjusting the ends  
50  
51 of LTR-pair candidates using the Smith–Waterman algorithm. These boundaries were re-adjusted  
52  
53 based on the occurrence of the following typical LTR-RE features: (a) the putative LTR-RE were  
54  
55 flanked by the dinucleotides TG and CA at 5' and 3' ends respectively; (b) a target-site duplication  
56  
57 (TSD) of 4–6 nt in length was present in the sequence; (c) a putative 15–18 nt primer binding site  
58  
59  
60  
61  
62  
63  
64  
65

1 (PBS) complementary to a tRNA at the end of the putative 5'-LTR was present in the sequence; and  
2  
3 (d) a 20–25-nt polypurine tract (PPT) just upstream of the 5' end of the 3' LTR was present in the  
4  
5 sequence. Putative LTR-REs were manually validated using DOTTER [52], verifying the occurrence  
6  
7 of LTRs, dinucleotides TG and CA at the 5' and 3' ends respectively, and TSDs. The validated LTR-  
8  
9 REs were annotated using BLASTX and BLASTN querying the NCBI nr nucleotide and protein  
10  
11 NCBI databases and RepBase [50]. To limit false-positive detection, a fixed E-value threshold of E  
12  
13  $< 10^{-5}$  for BLASTN and  $E < 10^{-10}$  for BLASTX was used. The full-length elements identified were  
14  
15 analysed using RepeatExplorer [49], performing searches for GAG, protease, retrotranscriptase,  
16  
17 RNaseH, integrase, and chromodomain derived from plant protein domains from RepBase. The  
18  
19 similarity search was filtered at E-value  $< 10^{-10}$ , allowing for both mismatches and frameshifts. The  
20  
21 same tool was used to assign full-length elements to specific *Gypsy* or *Copia* lineages. Full-length  
22  
23 LTR-REs that were identified as belonging to *Gypsy* or *Copia* superfamilies, and clusters annotated  
24  
25 as LTR-retrotransposons by RepeatExplorer (see above) were then used as reference datasets for  
26  
27 further searches in order to identify previously unclassified elements using RepeatMasker, running  
28  
29 default parameters, but with -div set to 20.  
30  
31  
32  
33  
34  
35  
36 For determination of RE redundancy, approximately 32,000,000 randomly-selected raw *Potentilla*  
37  
38 Illumina paired end reads, corresponding to 10.3× genome coverage. After removal of organellar  
39  
40 contamination performed by mapping the reads to an in-house Rosaceae organellar database and the  
41  
42 removal of duplicate reads, a total of 25,206,510 filtered nuclear reads corresponding to 7.2×  
43  
44 equivalent genomic coverage were used for redundancy analysis by mapping the reads to all REs  
45  
46 characterized in the *Potentilla* genome using CLC-BIO Genomic Workbench 8.0 (CLC-BIO, Aarhus,  
47  
48 Denmark). Mismatch cost, deletion cost, and insertion cost were fixed at 1, and similarity and length  
49  
50 fraction were both fixed at 0.9, 0.8, 0.5 or 0.4 to obtain high, medium, low, or very low stringencies,  
51  
52 respectively. As reads that mapped to multiple distinct sequences were few, and distributed randomly  
53  
54 throughout the dataset, the number of reads mapping to each RE was taken as the degree of  
55  
56 redundancy of that sequence within the genome. The effective abundance of a particular class of reads  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 was calculated as the proportion of the total number of reads mapped in each class, with respect to  
2 the overall number of genomic reads mapped, using optimal stringency parameters, i.e. where further  
3 relaxation of stringency did not significantly increase the number of mapped reads.

4  
5 The abundance of each single RE sequence in the genome was analysed by mapping *Potentilla* DNA  
6 reads, corresponding to 2× genome coverage to the full-length REs characterised, one by one using  
7 BWA (alignment via Burrows–Wheeler transformation) version 0.7.5a-r405 [53] running the  
8 following parameters: bwaaln -t 4 -l 12 -n 4 -k 2 -o 3 -e 3 -M 2 -O 6 -E 3. The resulting single-end  
9 mappings were resolved via the samse module of BWA, and the output was converted to .bam file  
10 format using SAMtools version 0.1.19 [54]. Subsequently, SAMtools was used to calculate the  
11 number of mapped reads for each alignment using the following parameters: samtools view -c -F 4.

## 12 **Determination of RE insertion age**

13 Retrotransposon insertion age was estimated through a sequence divergence comparison of the 5'-  
14 and 3'-LTRs of each putative full-length retrotransposon. Synonymous substitution rates were  
15 calculated for 50 pairs of orthologous gene sequences of *P. micrantha* and *F. vesca*, using a time of  
16 divergence of 24.22 million years [3]. Subsequently, the two LTRs were aligned with ClustalX  
17 software [53], indels were eliminated, and the number of nucleotide substitutions was counted using  
18 DnaSP [54] for each retrotransposon. The insertion times of retrotransposons with both LTRs were  
19 dated using the Kimura two parameter (K2P) method [55], calculated using DnaSP, and a  
20 synonymous substitution rate that is twofold that calculated for genes [56,57].

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

22 The data set supporting the results of this article are available in the GenBank repository, project  
23 number PRJEB18433. The genome reference sequence and gene predictions can be downloaded from  
24 the GigaScience GigaDB repository.

1 **FUNDING**

2 This work was funded by a grant to the Fondazione Edmund Mach (FEM) from the Autonomous  
3  
4 Province of Trento grants office. A.C. acknowledges funding from the Department of Agriculture,  
5  
6 Food and Environment of Pisa University, Project ‘Plantomics’.  
7  
8  
9

10  
11  
12 **CONFLICT OF INTERESTS**

13  
14  
15 The authors declare no competing interests.  
16  
17  
18  
19

20 **AUTHOR CONTRIBUTIONS**

21  
22 M.Buti performed the experiments, analysed and interpreted all data and authored the paper. M.M.,  
23  
24 P.S. and A.C. analysed sequence data and performed genome assemblies. K.E. and M. Brillì assisted  
25  
26 with experimental design, analysed and interpreted gene expression data and commented on and  
27  
28 contributed to the manuscript. L.N. and A.C. performed full-length retrotransposon isolation. E.B.,  
29  
30  
31 F.M. and A.C. performed clustering, annotation and redundancy analyses of repetitive sequences.  
32  
33  
34 E.B., F.M., L.N. and A.C. participated in the interpretation and discussion of results and contributed  
35  
36 to the writing of the paper. A.L and M.Borodovsky performed gene predictions and analysed and  
37  
38 interpreted the data. L.G., N.Š. assisted with experiments, interpreted data and contributed to the  
39  
40  
41 manuscript. M.A. and J.W. assisted with genome assemblies and gene annotation. C.V. analysed and  
42  
43 interpreted phylogenetic data and contributed to the manuscript. R.V. commented on the manuscript.  
44  
45  
46 D.J.S. designed the study, assisted with the experiments, analysed and interpreted the data and  
47  
48 authored the paper.  
49  
50  
51  
52  
53

54 **ADDITIONAL FILES**

55  
56 Additional File 1: Table S1. Illumina sequencing libraries used in the sequencing of the *Potentilla*  
57  
58 *micrantha* genome including fragment sizes and total genome depth of coverage.  
59  
60  
61  
62  
63  
64  
65

1 Additional File 2: Table S2. PacBio RS sequencing kits and chemistries used for *Potentilla micrantha*  
2 sequencing.  
3  
4 Additional File 3: Table S3. RNAseq read data used for gene prediction and number of splice sites  
5 identified in the *Potentilla micrantha* genome.  
6  
7 Additional File 4: Table S4. *Potentilla micrantha* and *Fragaria vesca* RNAseq reads statistics.  
8  
9 Additional File 5: Fig S1. Distribution of predicted genes *Potentilla micrantha* and *Fragaria vesca*  
10 mapped, blasted and GO-annotated by BLAST2GO analysis.  
11  
12 Additional File 6: Fig S2. The differential gene expression profiles between the four developmental  
13 stages of fruit development studied in *F. vesca* and *P. micrantha*.  
14  
15 Additional File 7: Fig S3. The overall abundance of different classes of transposons within the  
16 *Potentilla micrantha* genome according to the analyses performed using RepeatExplorer.  
17  
18 Additional File 8: Fig S4. Genome proportion in *Potentilla micrantha* and *Fragaria vesca* of 291  
19 repeats clustered using RepeatExplorer. Other repeats include satellite DNAs, pararetroviruses, and  
20 one LINE.  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35

## 36 REFERENCES

- 37 1. Eriksson T, Donoghue MJ, Hibbs MS. Phylogenetic analysis of *Potentilla* using DNA sequences  
38 of nuclear ribosomal internal transcribed spacers (ITS), and implications for the classification of  
39 Rosoideae (Rosaceae). *Plant Syst. Evol.* [Internet]. Springer-Verlag; 1998 [cited 2016 Aug  
40 9];211:155–79. Available from: <http://link.springer.com/10.1007/BF00985357>  
41  
42
- 43 2. Potter D, Eriksson T, Evans RC, Oh S, Smedmark JEE, Morgan DR, et al. Phylogeny and  
44 classification of Rosaceae. *Plant Syst. Evol.* [Internet]. 2007 [cited 2015 Oct 3];266:5–43. Available  
45 from: <http://link.springer.com/10.1007/s00606-007-0539-9>  
46  
47
- 48 3. Njuguna W, Liston A, Cronn R, Ashman T-L, Bassil N. Insights into phylogeny, sex function  
49 and age of *Fragaria* based on whole chloroplast genome sequencing. *Mol. Phylogenet. Evol.*  
50 2013;66:17–29.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1 11. Kang C, Darwish O, Geretz A, Shahan R, Alkharouf N, Liu Z. Genome-Scale Transcriptomic  
2 Insights into Early-Stage Fruit Development in Woodland Strawberry *Fragaria vesca*. *Plant Cell*  
3  
4 [Internet]. 2013;25:1960–78. Available from:  
5  
6 <http://www.plantcell.org/cgi/doi/10.1105/tpc.113.111732>  
7  
8
- 9 21. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome  
10 of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* [Internet]. 2011 [cited 2016 Aug 8];43:109–  
11  
12 16. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21186353>  
13  
14  
15  
16
- 17 4. Jung S, Cestaro A, Troglio M, Main D, Zheng P, Cho I, et al. Whole genome comparisons of  
18  
19 *Fragaria*, *Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies.  
20  
21 *BMC Genomics* [Internet]. 2012 [cited 2016 Aug 8];13:129. Available from:  
22  
23 <http://www.ncbi.nlm.nih.gov/pubmed/22475018>  
24  
25
- 26 5. Davidson RM, Gowda M, Moghe G, Lin H, Vaillancourt B, Shiu S-H, et al. Comparative  
27 transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.*  
28  
29 [Internet]. 2012 [cited 2016 Aug 8];71:492–502. Available from:  
30  
31  
32 <http://www.ncbi.nlm.nih.gov/pubmed/22443345>  
33  
34
- 35 6. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, et al. A genome  
36 triplication associated with early diversification of the core eudicots. *Genome Biol.* [Internet].  
37  
38 *BioMed Central*; 2012 [cited 2017 Feb 16];13:R3. Available from:  
39  
40  
41 <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-1-r3>  
42  
43
- 44 7. Ferrarini M, Moretto M, Ward JA, Šurbanovski N, Stevanović V, Giongo L, et al. An evaluation  
45 of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC*  
46  
47 *Genomics* [Internet]. *BioMed Central*; 2013 [cited 2016 Aug 8];14:670. Available from:  
48  
49  
50  
51 <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-670>  
52  
53
- 54 8. Dolezel J, Bartos J, Voglmayr H, Greilhuber J. Letter to the editor. *Cytometry* [Internet]. Wiley  
55  
56 *Subscription Services, Inc., A Wiley Company*; 2003 [cited 2016 Aug 9];51A:127–8. Available  
57  
58 from: <http://doi.wiley.com/10.1002/cyto.a.10013>  
59  
60  
61

- 1 9. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading  
2 Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. Liu Z, editor. PLoS  
3 One [Internet]. Public Library of Science; 2012 [cited 2016 Aug 8];7:e47768. Available from:  
4  
5  
6  
7 4 <http://dx.plos.org/10.1371/journal.pone.0047768>  
8
- 9 10. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: assessing  
10 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics  
11  
12 6 [Internet]. 2015 [cited 2017 Nov 2];31:3210–2. Available from:  
13  
14 7 <http://www.ncbi.nlm.nih.gov/pubmed/26059717>  
15  
16
- 17 8 12. Day RC, Herridge RP, Ambrose BA, Macknight RC. Transcriptome Analysis of Proliferating  
18 Arabidopsis Endosperm Reveals Biological Implications for the Control of Syncytial Division,  
19 Cytokinin Signaling, and Gene Expression Regulation. PLANT Physiol. [Internet]. American  
20 Society of Plant Biologists; 2008 [cited 2016 Aug 10];148:1964–84. Available from:  
21  
22 10 <http://www.plantphysiol.org/cgi/doi/10.1104/pp.108.128108>  
23  
24 11
- 25 13. Hehenberger E, Kradolfer D, Köhler C. Endosperm cellularization defines an important  
26 developmental transition for embryo development. Development [Internet]. 2012 [cited 2016 Aug  
27 10];139:2031–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22535409>  
28  
29 13
- 30 14. Fang S-C, Fernandez DE. Effect of regulated overexpression of the MADS domain factor  
31 AGL15 on flower senescence and fruit maturation. Plant Physiol. [Internet]. 2002 [cited 2016 Aug  
32 10];130:78–89. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12226488>  
33  
34 15
- 35 15. VanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, et al. The genome of  
36 black raspberry (*Rubus occidentalis*). Plant J. [Internet]. 2016 [cited 2016 Aug 16]; Available from:  
37  
38  
39 17 <http://www.ncbi.nlm.nih.gov/pubmed/27228578>  
40  
41 16
- 42 16. Dickson EE, Arumuganathan K, Kresovich S, Doyle JJ, Kresovich S, Doyle2 JJ. Nuclear DNA  
43 Content Variation within the Rosaceae NUCLEAR DNA CONTENT VARIATION WITHIN THE  
44 ROSACEAE'. Am. J. Bot. Am. J. Bot. Am. J. Bot. [Internet]. 1992 [cited 2016 Nov 5];79:1081–6.  
45  
46  
47  
48  
49 21 Available from: [http://scholarcommons.sc.edu/biol\\_facpub](http://scholarcommons.sc.edu/biol_facpub)  
50  
51 22  
52  
53 23  
54  
55  
56 24  
57  
58 25  
59  
60  
61 26

- 1 17. Meng R, Finn C. Determining Ploidy Level and Nuclear DNA Content in Rubus by Flow  
1  
2 2 Cytometry. *J. Am. Soc. Hortic. Sci. American Society for Horticultural Science*; 2002;127:767–75.  
3  
4  
5 3 18. Rajapakse S, Byrne DH, Zhang L, Anderson N, Arumuganathan K, Ballard RE. Two genetic  
6  
7 4 linkage maps of tetraploid roses. *TAG Theor. Appl. Genet.* [Internet]. Springer-Verlag; 2001 [cited  
8  
9  
10 5 2016 Nov 5];103:575–83. Available from: <http://link.springer.com/10.1007/PL00002912>  
11  
12 6 19. Yokoya K, Roberts A V., Mottley J, Lewis R, Brandham PE. Nuclear DNA Amounts in Roses.  
13  
14 7 *Ann. Bot.* [Internet]. Oxford University Press; 2000 [cited 2016 Nov 5];85:557–61. Available from:  
15  
16  
17 8 <http://aob.oxfordjournals.org/cgi/doi/10.1006/anbo.1999.1102>  
18  
19 9 20. Vitte C, Fustier M-A, Alix K, Tenaillon MI. The bright side of transposons in crop evolution.  
20  
21  
22 10 *Brief. Funct. Genomics* [Internet]. Oxford University Press; 2014 [cited 2016 Aug 15];13:276–95.  
23  
24 11 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24681749>  
25  
26  
27 12 22. Suzuki M, Kamide Y, Nagata N, Seki H, Ohyama K, Kato H, et al. Loss of function of 3-  
28  
29 13 hydroxy-3-methylglutaryl coenzyme A reductase 1 (HMG1) in Arabidopsis leads to dwarfing, early  
30  
31 14 senescence and male sterility, and reduced sterol levels. *Plant J.* [Internet]. 2004 [cited 2017 Nov  
32  
33  
34 15 2];37:750–61. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/14871314>  
35  
36  
37 16 23. Schrick K, Debolt S, Bulone V. Deciphering the molecular functions of sterols in cellulose  
38  
39 17 biosynthesis. *Front. Plant Sci.* [Internet]. Frontiers Media SA; 2012 [cited 2017 Nov 2];3:84.  
40  
41 18 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22639668>  
42  
43  
44 19 24. Smaczniak C, Immink RGH, Angenent GC, Kaufmann K, Adamczyk BJ, Fernandez DE, et al.  
45  
46 20 Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent  
47  
48 21 studies. *Development* [Internet]. Oxford University Press for The Company of Biologists Limited;  
49  
50  
51 22 2012 [cited 2016 Aug 15];139:3081–98. Available from:  
52  
53  
54 23 <http://www.ncbi.nlm.nih.gov/pubmed/22872082>  
55  
56 24 25. Shirzadi R, Andersen ED, Bjerkan KN, Gloeckle BM, Heese M, Ungru A, et al. Genome-wide  
57  
58 25 transcript profiling of endosperm without paternal contribution identifies parent-of-origin-  
59  
60  
61 26 dependent regulation of AGAMOUS-LIKE36. *PLoS Genet.* [Internet]. 2011 [cited 2016 Aug  
62  
63  
64  
65



- 1 16];7:e1001303. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21379330>
- 2 26. Harding EW, Tang W, Nichols KW, Fernandez DE, Perry SE. Expression and maintenance of  
3 embryogenic potential is enhanced through constitutive expression of AGAMOUS-Like 15. *Plant*  
4  
5 3 *Physiol.* [Internet]. 2003 [cited 2016 Aug 16];133:653–63. Available from:  
6  
7 4 <http://www.ncbi.nlm.nih.gov/pubmed/14512519>
- 8  
9  
10 5  
11  
12 6 27. Serivichyaswat P, Ryu H-S, Kim W, Kim S, Chung KS, Kim JJ, et al. Expression of the floral  
13  
14 7 repressor miRNA156 is positively regulated by the AGAMOUS-like proteins AGL15 and AGL18.  
15  
16 *Mol. Cells* [Internet]. Korean Society for Molecular and Cellular Biology; 2015 [cited 2016 Aug  
17 8  
18 16];38:259–66. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25666346>
- 19 9  
20  
21  
22 10 28. Chen D-H, Ronald PC. A Rapid DNA Miniprep Method Suitable for AFLP and Other  
23  
24 11 PCR Applications. *Plant Mol. Biol. Report.* [Internet]. Kluwer Academic Publishers; 1999 [cited  
25  
26 12 2016 Aug 8];17:53–7. Available from: <http://link.springer.com/10.1023/A:1007585532036>
- 27  
28  
29 13 29. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, et al. ALLPATHS:  
30  
31 14 de novo assembly of whole-genome shotgun microreads. *Genome Res.* [Internet]. 2008 [cited 2016  
32  
33 Aug 8];18:810–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18340039>
- 34 15  
35  
36 16 30. Smit AFA, Hubley R. RepeatModeler - 1.0.7 [Internet]. 2013. Available from:  
37  
38 <http://www.repeatmasker.org/RepeatModeler.html>
- 39 17  
40  
41 18 31. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. 2013. Available from:  
42  
43 <http://www.repeatmasker.org/>
- 44 19  
45  
46 20 32. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic  
47  
48 21 training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* [Internet]. Oxford University  
49  
50 Press; 2014 [cited 2016 Aug 8];42:e119. Available from:  
51 22  
52 <http://www.ncbi.nlm.nih.gov/pubmed/24990371>
- 53 23  
54  
55  
56 24 33. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast  
57  
58 25 universal RNA-seq aligner. *Bioinformatics* [Internet]. Oxford University Press; 2013 [cited 2016  
59  
60 Aug 8];29:15–21. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23104886>
- 61 26  
62  
63  
64  
65

- 1 34. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics.  
2 Int. J. Plant Genomics [Internet]. Hindawi Publishing Corporation; 2008 [cited 2016 Aug  
3  
4 8];2008:619832. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18483572>  
5  
6  
7 35. Ostlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. InParanoid 7: new  
8  
9 algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. [Internet]. 2010 [cited  
10 5  
11 2016 Aug 10];38:D196-203. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19892828>  
12 6  
13  
14 36. Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, et al. GDR (Genome Database for  
15 7  
16 Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic Acids Res.  
17 8  
18 [Internet]. Oxford University Press; 2008 [cited 2016 Aug 9];36:D1034-40. Available from:  
19 9  
20  
21 <http://www.ncbi.nlm.nih.gov/pubmed/17932055>  
22 10  
23  
24 37. Andrews S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput  
25 4  
26 Sequence Data [Internet]. 2010. Available from:  
27 12  
28  
29 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>  
30 13  
31  
32 38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.  
33 14  
34 Bioinformatics [Internet]. Oxford University Press; 2014 [cited 2016 Aug 9];30:2114–20. Available  
35 15  
36 from: <http://www.ncbi.nlm.nih.gov/pubmed/24695404>  
37 16  
38  
39 39. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods [Internet].  
40 17  
41 NIH Public Access; 2012 [cited 2016 Aug 8];9:357–9. Available from:  
42 18  
43  
44 <http://www.ncbi.nlm.nih.gov/pubmed/22388286>  
45 19  
46  
47 40. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying  
48 20  
49 mammalian transcriptomes by RNA-Seq. Nat. Methods [Internet]. Nature Publishing Group; 2008  
50 21  
51 [cited 2016 Aug 8];5:621–8. Available from: <http://www.nature.com/doifinder/10.1038/nmeth.1226>  
52 22  
53  
54 41. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int. J.  
55 23  
56 Plant Genomics [Internet]. Hindawi Publishing Corporation; 2008 [cited 2016 Aug 8];2008:619832.  
57 24  
58 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18483572>  
59 25  
60  
61  
62  
63  
64  
65

- 1 42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software  
2 environment for integrated models of biomolecular interaction networks. *Genome Res.* [Internet]. Cold  
3 Spring Harbor Laboratory Press; 2003 [cited 2017 Nov 3];13:2498–504. Available from:  
4 <http://www.ncbi.nlm.nih.gov/pubmed/14597658>  
5  
6  
7  
8  
9 43. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD:  
10 NCBI's conserved domain database. *Nucleic Acids Res.* [Internet]. 2015 [cited 2016 Aug  
11 8];43:D222-6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25414356>  
12  
13  
14 44. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features.  
15 *Bioinformatics* [Internet]. 2010 [cited 2016 Aug 8];26:841–2. Available from:  
16  
17  
18  
19 <http://www.ncbi.nlm.nih.gov/pubmed/20110278>  
20  
21  
22  
23  
24 45. Liu K, Warnow TJ, Holder MT, Nelesen SM, Yu J, Stamatakis AP, et al. SATE-II: very fast and  
25 accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Syst.*  
26 *Biol.* [Internet]. Oxford University Press; 2012 [cited 2016 Aug 9];61:90–106. Available from:  
27  
28  
29  
30  
31 <http://www.ncbi.nlm.nih.gov/pubmed/22139466>  
32  
33  
34 46. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version  
35 7.0 for Bigger Datasets. *Mol. Biol. Evol.* [Internet]. 2016;33:1870–4. Available from:  
36  
37  
38  
39 <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw054>  
40  
41 47. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic  
42 trees made easy. *Nucleic Acids Res.* [Internet]. Oxford University Press; 2011 [cited 2016 Aug  
43 8];39:W475-8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21470960>  
44  
45  
46  
47  
48 48. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, et al. The  
49 *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information  
50 retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* [Internet]. Oxford  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

- 1 server for genome-wide characterization of eukaryotic repetitive elements from next-generation  
1  
2 2 sequence reads. *Bioinformatics* [Internet]. 2013 [cited 2016 Aug 9];29:792–3. Available from:  
3  
4  
5 3 <http://www.ncbi.nlm.nih.gov/pubmed/23376349>  
6
- 7 4 50. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update,  
8  
9  
10 5 a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* [Internet]. Karger Publishers;  
11  
12 6 2005 [cited 2016 Aug 9];110:462–7. Available from:  
13  
14 7 <http://www.karger.com/?doi=10.1159/000084979>  
15  
16
- 17 8 51. Xu Z, Wang H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR  
18  
19 9 retrotransposons. *Nucleic Acids Res.* [Internet]. Oxford University Press; 2007 [cited 2016 Aug  
20  
21  
22 10 8];35:W265-8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17485477>  
23
- 24 11 52. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for  
25  
26  
27 12 genomic DNA and protein sequence analysis. *Gene* [Internet]. 1995 [cited 2016 Aug 8];167:GC1-  
28  
29 13 10. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8566757>  
30
- 31 14 53. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.  
32  
33  
34 15 *Bioinformatics* [Internet]. 2009 [cited 2016 Aug 9];25:1754–60. Available from:  
35  
36 16 <http://www.ncbi.nlm.nih.gov/pubmed/19451168>  
37  
38
- 39 17 54. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence  
40  
41 18 Alignment/Map format and SAMtools. *Bioinformatics* [Internet]. 2009 [cited 2016 Aug  
42  
43  
44 19 9];25:2078–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19505943>  
45
- 46 20 55. Kimura M. A simple method for estimating evolutionary rates of base substitutions through  
47  
48  
49 21 comparative studies of nucleotide sequences. *J. Mol. Evol.* [Internet]. 1980 [cited 2016 Aug  
50  
51 22 9];16:111–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/7463489>  
52
- 53 23 56. Sanmiguel P, Bennetzen JL. Evidence that a Recent Increase in Maize Genome Size was  
54  
55  
56 24 Caused by the Massive Amplification of Intergene Retrotransposons. *Ann. Bot.* Oxford University  
57  
58 25 Press; 1998;82:37–44.  
59
- 60  
61 26 57. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl.*  
62  
63  
64  
65

1 Acad. Sci. U. S. A. [Internet]. National Academy of Sciences; 2004 [cited 2016 Aug 9];101:12404–  
2 10. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15240870>

## 7 4 **FIGURE LEGENDS AND TABLES**

9 5 **Figure 1.** Comparison of *Fragaria vesca* and *Potentilla micrantha* morphology for leaves, flowers  
11 and fruits.

14 7 **Figure 2a.** Anchoring of five *Potentilla micrantha* genome scaffolds to the *Fragaria vesca* Fvb  
16 pseudomolecules *Fvb2* and *Fvb4* demonstrating the microsynteny between the *F. vesca* and *P.*  
18 *micrantha* genomes (numbers in parentheses below the scaffold names indicate the number of genes  
19 9 contained in each split syntenic block.

24 11 **Figure 2b.** A comparison of the seven pseudomolecules of the *F. vesca* genome with eight *P.*  
25 *micrantha* sequencing scaffolds, highlighting the major translocation events identified between the  
26 12 two species in this investigation.

31 14 **Figure 3.** *Potentilla micrantha* flower/fruit developmental stages used for RNA extraction.

34 15 **Figure 4.** Differentially expressed genes during fruit development in *P. micrantha* and *F. vesca*.

36 16 Volcano plots of differential expression analysis between the four developmental stages A-B-C-D in  
37 *Potentilla micrantha* and *Fragaria vesca*. Using a cut-off of  $\sqrt{\text{MSR}} > 2.00$  and  $p\text{-value} < 10^{-3}$ ,  
38 17 1,556 genes were differentially expressed in *Potentilla micrantha*, whilst 816 genes were  
40 differentially expressed in *Fragaria vesca*.

46 20 **Figure 5.** Over-represented GO-slim categories in *Fragaria vesca* and *Potentilla micrantha* DEGs  
47 sets. The circles are shaded based on significance level (yellow = FDR below 0.05), and the radius of  
48 21 each circle is proportional to the number of genes included in each GO-slim category.

53 23 **Figure 6.** Heatmap comparing the log expression values of 205 genes (orthologs of both *F. vesca* and  
54 *P. micrantha*) The rows (genes) were sorted using hierarchical clustering using 'correlation' distance  
55 and 'complete' linkage. A-D correspond to the four developmental stages defined in the methods  
56 24 section.

**Figure 7.** A Maximum Likelihood-based phylogenetic reconstruction of the *Potentilla micrantha* and *Fragaria vesca* genes containing MADS-box motifs, along with the relative gene expression levels for each gene. Categories A-D refer to the developmental stages defined in the methods. Filled circles represent the relative level of support for each relationship defined in the Maximum Likelihood analysis.

**Figure 8.** The three identified clades of orthologous MADS-box motif containing genes that were not expressed or poorly expressed in *Potentilla micrantha* but highly expressed in *Fragaria vesca*. Categories A-D refer to the four developmental stages defined in the methods.

**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

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

Superfamily	Family	Number	Percentage
Ty1-Copia	<i>AleI/Retrofit</i>	14	2.77
	<i>AleII</i>	26	5.15
		33	

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 1  
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

<i>Angela</i>	20	3.96
<i>Bianca</i>	114	22.57
<i>Ivana</i>	23	4.55
<i>Maximus/SIRE</i>	10	1.98
<i>TAR/Tork</i>	11	2.18
Unknown	2	0.40

---

Total	220	43.56
-------	-----	-------

---

Ty3-Gypsy	<i>Athila</i>	3	0.59
	<i>Chromovirus</i>	42	8.32
	<i>Ogre/TAT</i>	186	36.83
	Unknown	25	4.95
	Total	256	50.69

---

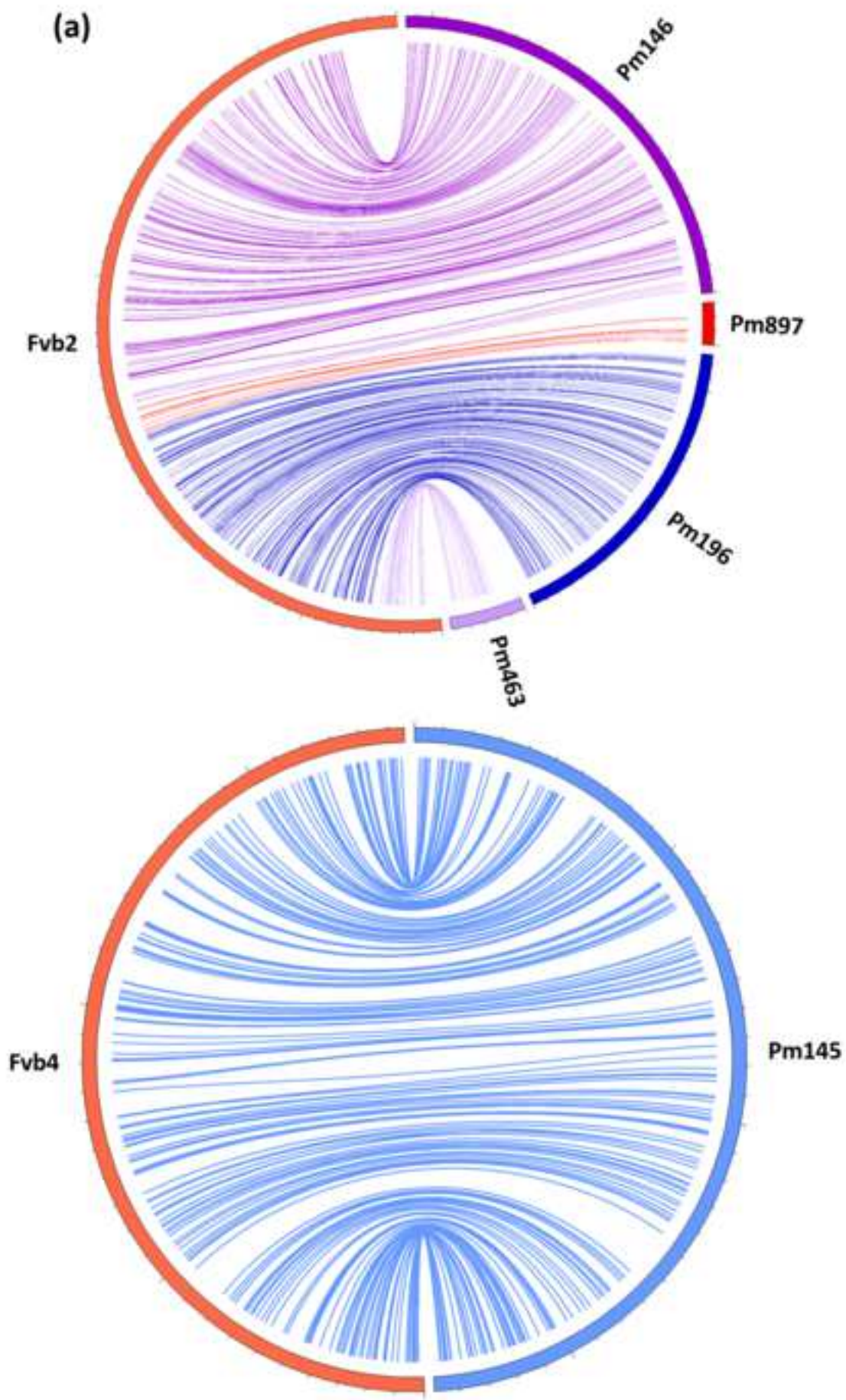
Unclassified		29	5.74
--------------	--	----	------

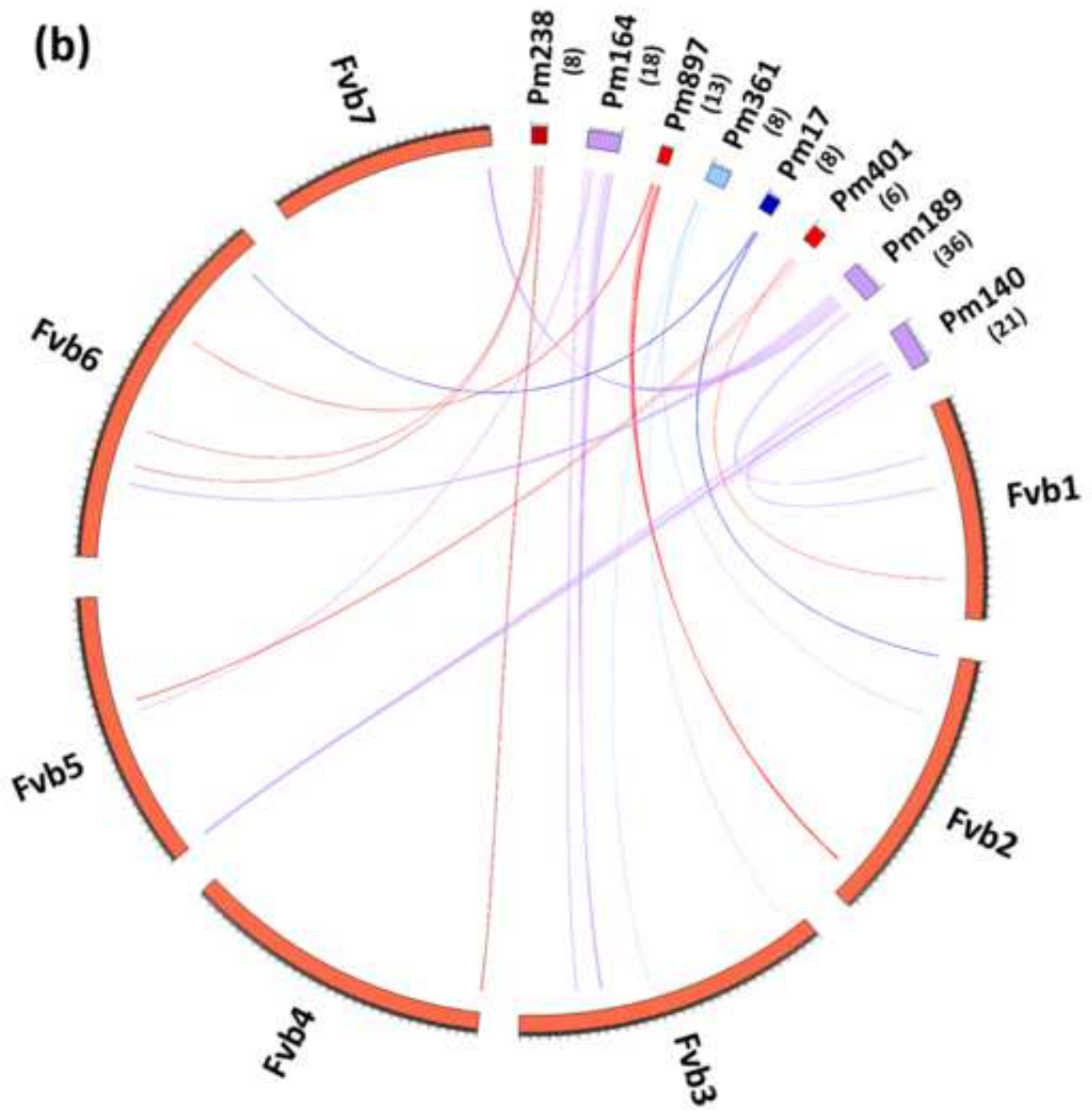
---



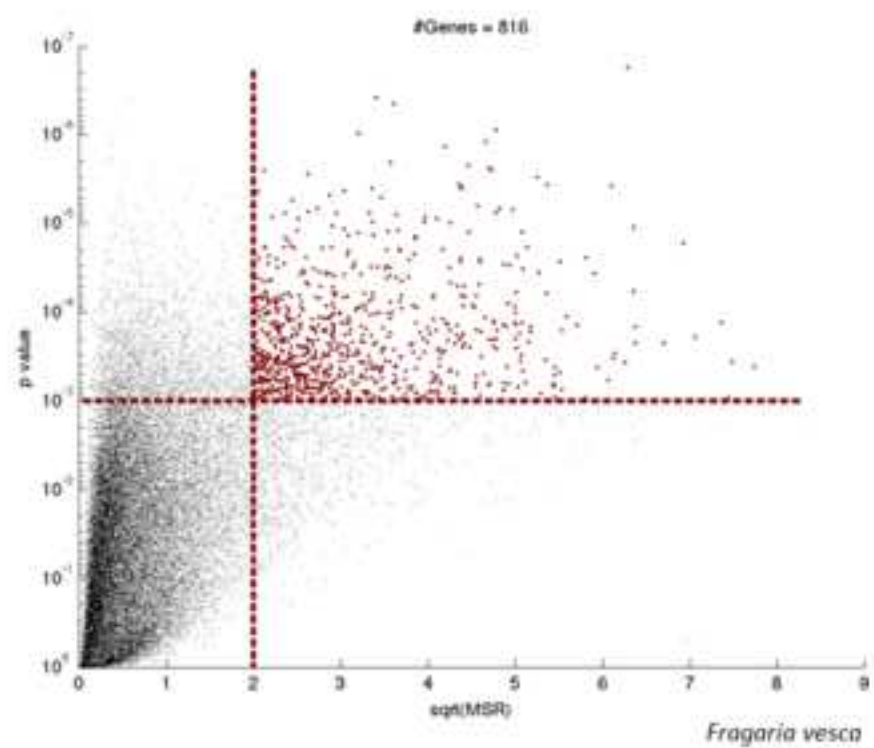
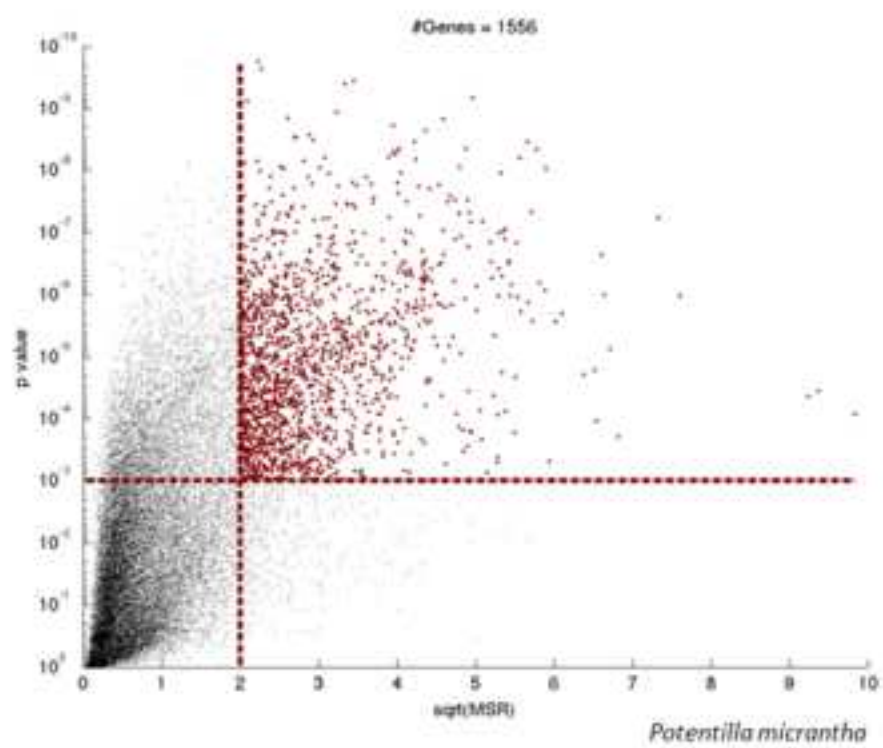




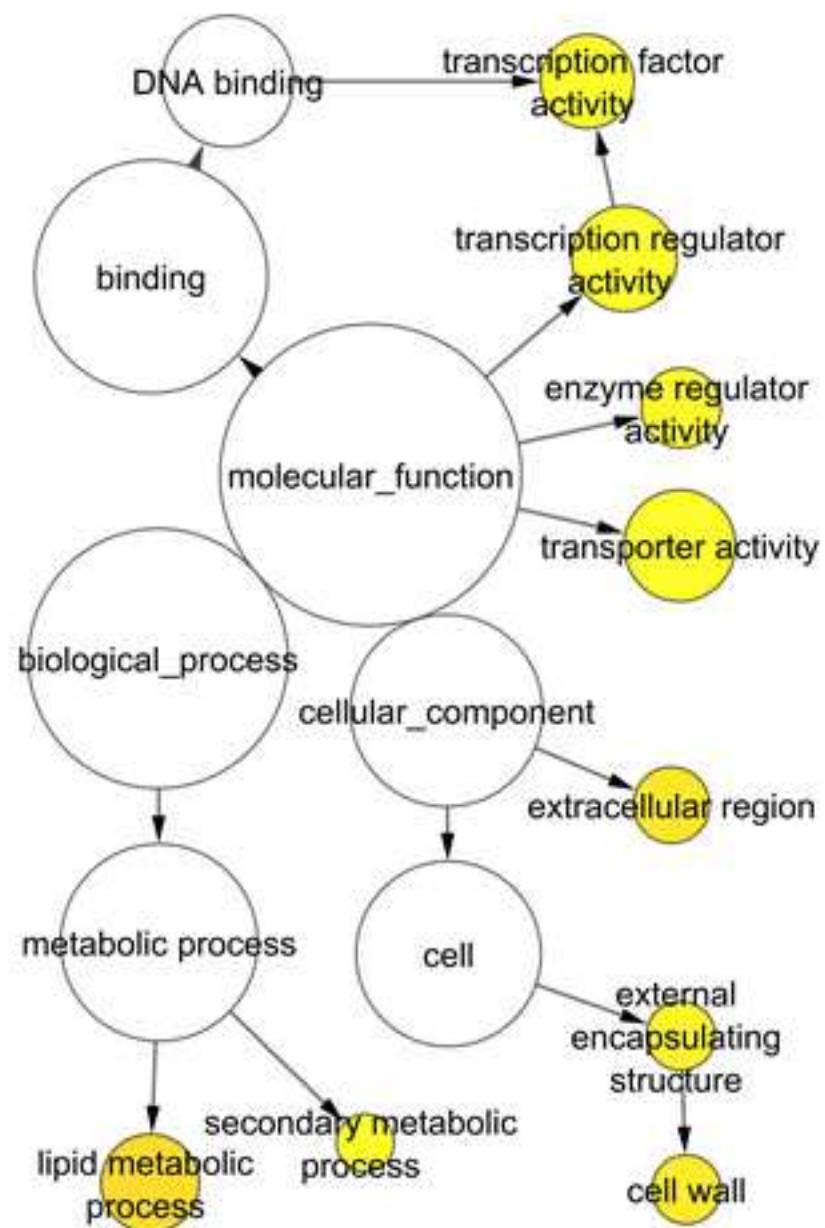
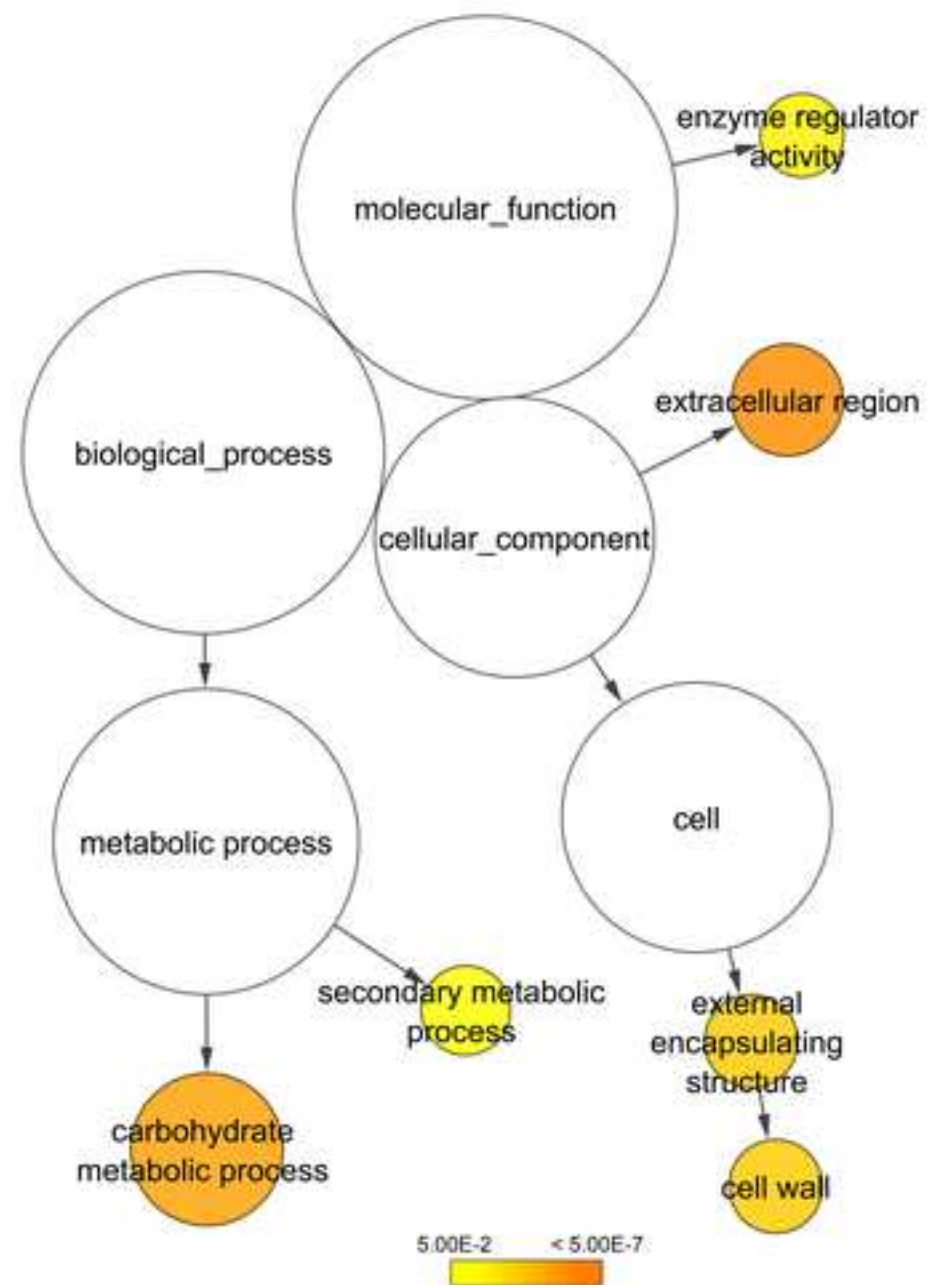










*Fragaria vesca**Potentilla micrantha*

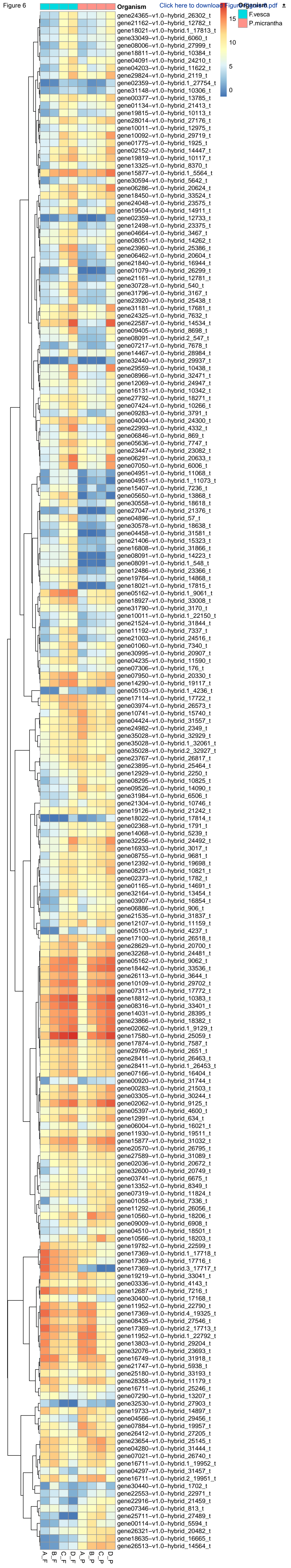
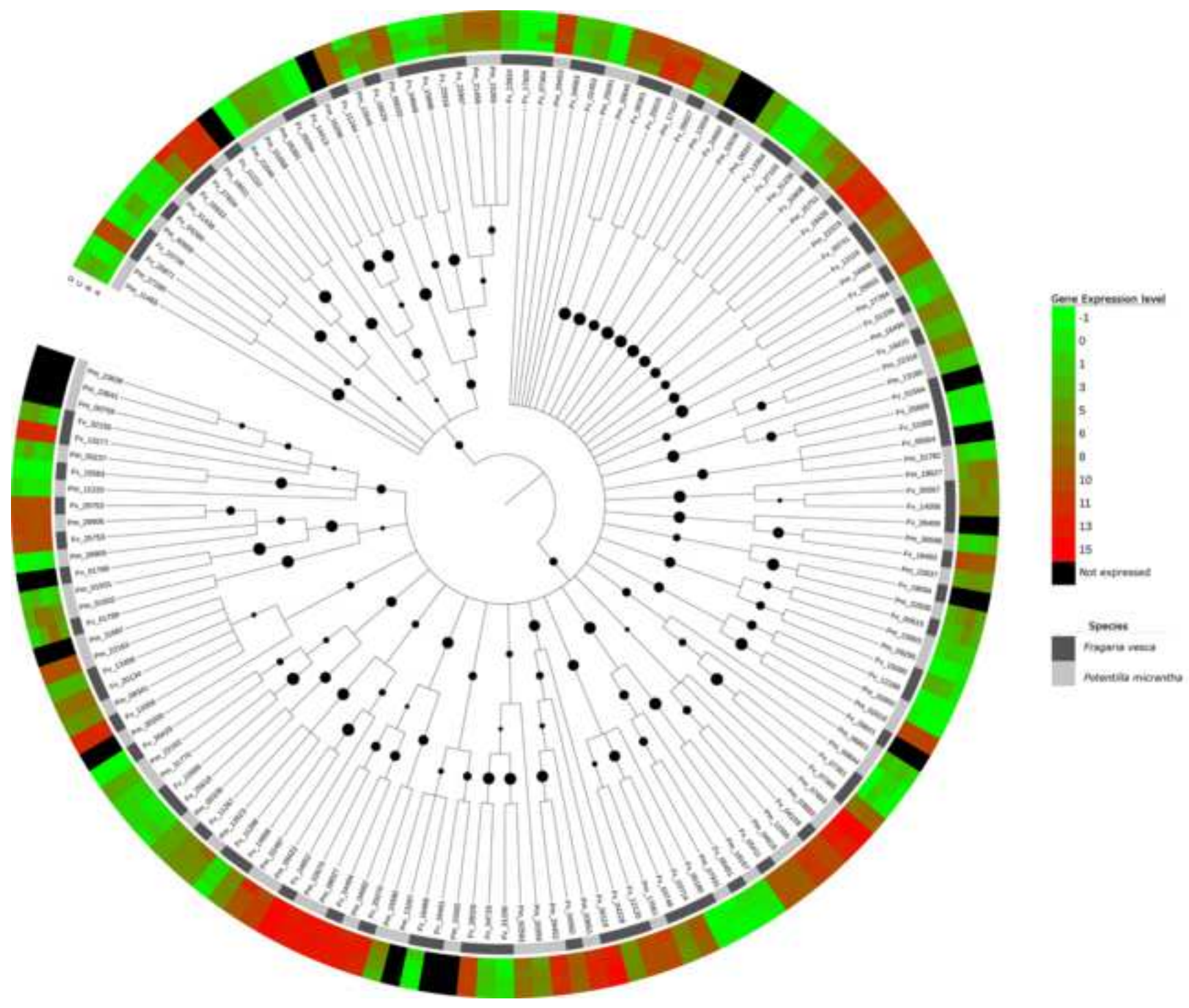
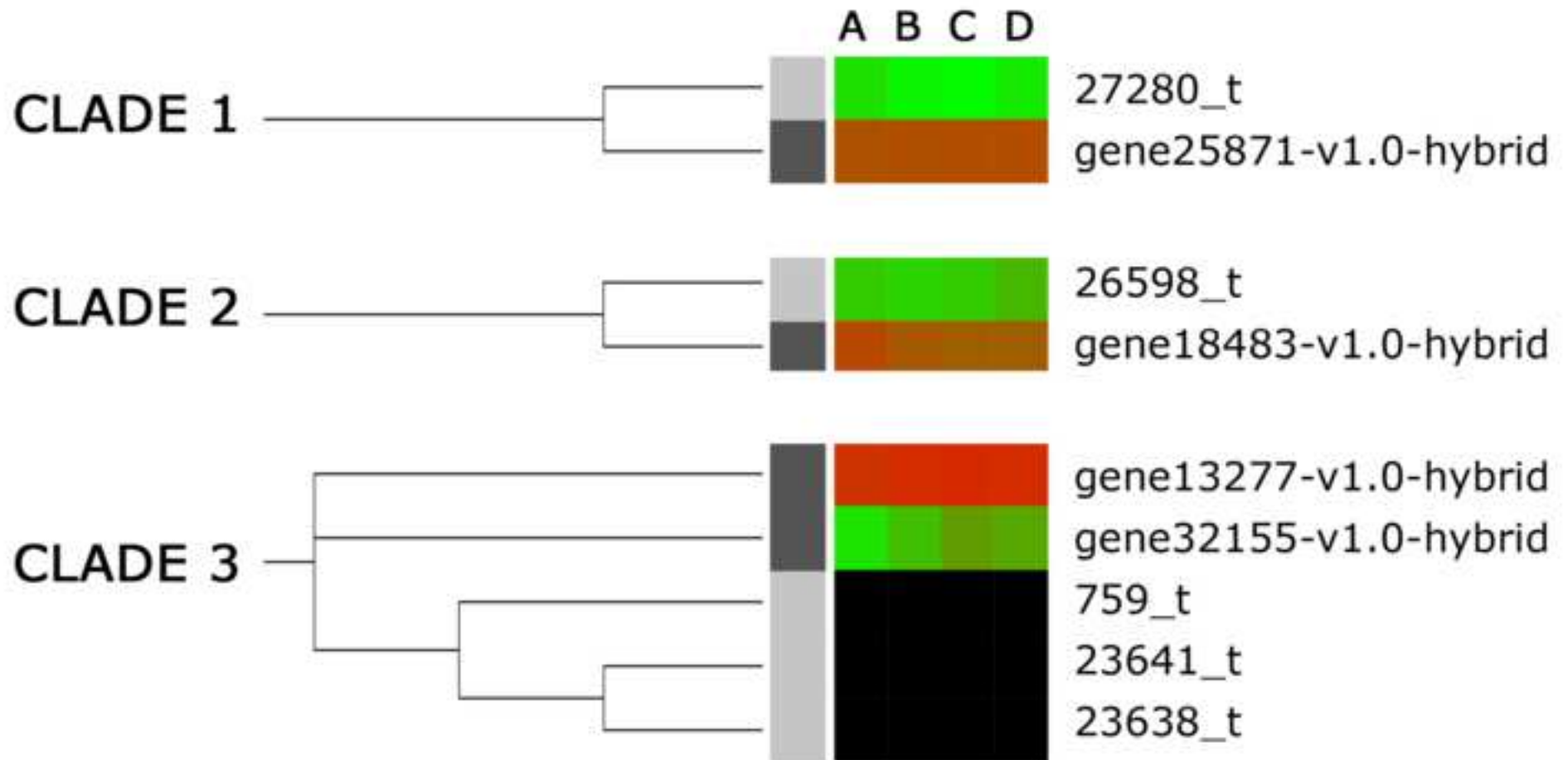




Figure 7

[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



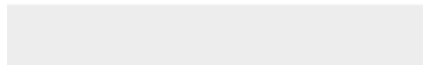


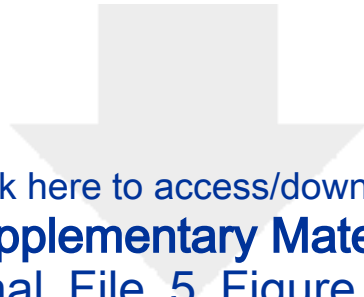
Click here to access/download  
**Supplementary Material**  
Additional\_File\_3\_Table S3.docx





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



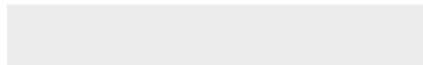


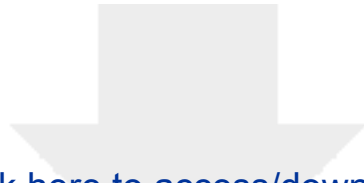
Click here to access/download  
**Supplementary Material**  
Additional\_File\_6\_Figure\_S2.png





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

