RESEARCH

CHILEANJAR
CHILEAN JOURNAL of AGRICULTURAL RESEARCH

Transcriptome assembly and expression analysis in *Colletotrichum gloeosporioides*-tolerant *Rubus glaucus* Benth.

Juliana Arias[1], Juan C. Rincón[1*], Ana M. López[1], and Marta L. Marulanda[1]

[1]Universidad Tecnológica de Pereira, Grupo de Investigación en Biodiversidad y Biotecnología, Carrera 27 No. 10-02 Pereira, Risaralda, Colombia. [*]Corresponding author (rincon.juan@utp.edu.co).

Received: 9 April 2019; Accepted: 30 July 2019; doi:10.4067/S0718-58392019000400565

ABSTRACT

Andean blackberry (*Rubus glaucus* Benth.) is an important crop of the Andean region affected by *Colletotrichum gloeosporioides*. In Colombia, tolerant plant material has been detected, but it has not been completely characterized. The objective of this research was oriented to analyze *de novo* transcriptome assembly of *R. glaucus*, and the comparison of the assembly with different reference genomes to further complete differential expression analysis of *R. glaucus* tolerant to *C. gloespoiorides* attack. To achieve this, three groups were used: infected tolerant material, infected susceptible material, and a susceptible group without inoculation. The RNA-seq sequencing was achieved through Illumina Hi-seq 2000. *De novo* assembly (Trinity, CD-HIT, TopHat) and functional annotation of sequences were carried out, additionally, mapping with reference genomes belonging to Rosaceae families was conducted (Bowtie2, TopHat). Subsequently, the differential expression was quantified (Cuffdiff) and analyzed through EdgeR. Variant analysis was made using MISA and SAMtools. After editing and assembly, 43579 consensus sequences were obtained (N50 = 489 bp; GC = 44.6%), annotation detected 35824 and 35602 sequences in Nt (partially non-redundant nucleotide sequences) and Nr (non-redundant protein sequences) databases, respectively. The 85% of Nr sequences was linked to members of Rosaceae family, mainly strawberry (67.6%). A total of 3570 simple sequence repeat (SSR) markers and 38791 single nucleotide polymorphisms (SNP) were found. The transcriptome of tolerant plants exhibited less SNPs. Finally, differentially expressed genes were found, including *RPM1*, *MAPKBP1*, *CKX2*. This research represents a contribution for future understanding of *R. glaucus* transcriptome, since there is little information available, and it would help further tolerance-related analysis.

Key words: Blackberry, fungal tolerance, molecular markers, RNA-seq, tolerance-related genes.

INTRODUCTION

Rosaceae family possess a great economic importance and wide distribution, including crops with economic relevance such as strawberry, pear, cherries, peach, raspberry, rose, among others. Several members of this family inhabit America and some of them are largely planted in Colombia, as is the case of Andean blackberry (*Rubus glaucus* Benth.) where different cultivars are found and, commonly, vegetative propagation is carried out.

Rubus glaucus culture is an important economic activity in Colombia and the Andean region, with a growing global market due to the demonstrated beneficial effects of its polyphenolic compounds in humans. A huge number of families from these regions depend on this agricultural activity, notwithstanding, the productive performance does not fulfill expectations due to economic losses generated by some diseases, among which anthracnose caused by *Colletotrichum gloeosporioides* (Penz.) Penz. & Sacc., outstands (López-Vásquez et al., 2013). Colombian Coffee Producing Region (CCPR) has reported incidences around 52.9% (Botero et al., 2002), making *C. gloeosporioides* the most important disease for *R. glaucus*.

Previous studies have indicated some *Colletotrichum* species as the main etiologic agents of anthracnose in the CCPR, where *C. gloeosporioides* caused 81% of the reported cases. These findings allowed to standardize the isolation method and plant inoculation techniques (Marulanda et al., 2007; 2014), making it possible to further identify and characterize plant material tolerant to this pathogen (López-Vásquez et al., 2013). In addition, the molecular characterization of several Andean blackberry varieties, including some tolerant accessions (Marulanda et al., 2012). Nevertheless, the knowledge towards genetics, genome and transcriptomics of the plant is still limited because the majority of these approaches have focused on other Rosaceae family species (Genome Database for Rosaceae https://www.rosaceae.org). This situation creates the necessity of developing experimental work oriented towards the elucidation of genetic composition, which allows the future identification of variants associated to productive traits, such as fruit quality and disease tolerance (including *C. gloeosporioides* tolerance), and consider them in propagation and breeding schemes.

Inside the *Rubus* genus various genes associated to fruit quality, metabolism of propanoids, and disease resistance (Zheng and Hrazdina, 2010; Han et al., 2017) have been described, almost all those genes evaluated over raspberry. Recently, Garcia-Seco et al. (2015) completed the assembly and transcriptome analysis of *Rubus* sp. var. Loch Ness, a blackberry species different to the commercially cultivated Andean blackberry, considering the fruit only and ignoring the consequences of *C. gloeosporioides* inoculation. Moreover, new versions of reference genomes of other Rosaceae species have been released (Edger et al., 2018; Saint-Oyant et al., 2018; VanBuren et al., 2018), which can be used in expression analysis. It makes it necessary to develop *R. glaucus* transcriptome analysis that allow us to understand its genetic peculiarity and improve the understanding of transcriptome and expression of tolerance-related genes.

Finally, it is important to mention that RNA-seq is a tool for transcriptome analysis and can be used in differential expression analysis under diverse circumstances, but variable results can be obtained whether reference genomes already exist or a *de novo* analysis is developed (Trapnell et al., 2010; Garcia-Seco et al., 2015). Based on the above mentioned information, the aim of this research was to analyze the transcriptome assembly of *R. glaucus* employing *de novo* assembly and with different reference genomes from Rosaceae species, to further complete differential expression analysis of Andean *R. glaucus* tolerant to *C. gloeosporioides*.

MATERIALS AND METHODS

Plant material and isolation of RNA

The plant material sampled for this study corresponded to commercial cultivars located in the municipality of Guática (5°20'26.2" N, 75°47'28" W; 2160 m a.s.l.), Risaralda, Colombia, with an accumulated precipitation of 1588.7 mm yr^{-1} (López-Vásquez et al., 2013). Sampled plants were propagated *in vitro* at the Laboratory of Plant Biotechnology of Universidad Tecnológica de Pereira. For the completion of this study, previously described *Colletotrichum gloeosporioides*-tolerant varieties (López-Vásquez et al., 2013), and susceptible material were used. Both were identified by its genetic and morphological traits and characterized employing molecular markers (López-Vásquez et al., 2013). Inoculation was accomplished on these sample groups (susceptible and tolerant) with an inoculum concentration of 1.2×10^6 spores mL^{-1} and incubated for 72 h. Infection was performed employing two highly pathogenic *C. gloeosporioides* strains (3S1 and 6) previously characterized (Marulanda et al., 2007).

At this section, three sample groups of plant material were established: the first consisted in *C. gloeosporioides*-susceptible with no inoculation (CSNI) to which sterile water was applied; the second corresponded to *C. gloeosporioides*-susceptible inoculated (CSI); and the third considered *C. gloeosporioides*-tolerant inoculated (CTI). After inoculation, a 72 h leaf sampling in liquid nitrogen was accomplished and immediately RNA extraction was carried out employing the FastTrack MAG mRNA isolation kit (Thermo Fisher Scientific, Waltham, Massachusetts, USA). Samples were treated with DNases for its further cDNA synthesis and sequencing.

Sequencing

Once RNA was extracted, library preparation and sequencing was carried out employing Illumina HiSeqTM 2000 technology (Beijing Genomics Institute BGI, Hong-Kong, China). Genomic library was constructed in paired-end mode following Illumina TruSeq RNA protocol. Reads were further processed considering adapter trimming and short (< 36 bp) low quality sequence suppressing (Phred < 20) with Trimmomatic software v0.38 (Bolger et al., 2014).

Transcriptome assembly

A revision in the Genome Database for Rosaceae (GDR; https://www.rosaceae.org) was made in order to assemble the obtained sequences. Different reference genomes of Rosaceae family species, including an assembly for *Rubus* genus in initial stages were found, for this reason it was decided to employ different assembly scenarios, first using reported reference genomes and last, a *de novo* scenario. In *de novo* assembly, the RNA-seq sequences of the three treatments were gathered and assembled in transcripts using Trinity v.2.5.1 software, which constructs Bruijn graphs from grouped sequences aiming to find the way that allows to report contig length, identify isoforms and paralog genes (Grabherr et al., 2011). Afterwards, reads of every treatment were aligned and mapped with Bowtie 2 v.2.3.2 (Langmean and Salzberg, 2012) and TopHat v2.1.1 (Trapnell et al., 2009). In order to reduce redundancy, sequence clusters with 90% of identity were formed through CD-HIT v4.6.8 (Fu et al., 2012). Unique sequences of each cluster were defined as unigenes.

Unigenes were aligned in BlastX against Non Redundant National Center for Biotechnology Information database (Nr NCBI), using an E-value cutoff of 0.000001 to evaluate taxonomic assignation. Subsequently, queries were made in Nt (Partially non-redundant nucleotide sequences) database, Swiss-Prot and functional annotation with BLAST2GO (Conesa and Götz, 2008) in database KEGG, Clusters of Orthologous Groups (COG) and Gene Ontology (GO). Annotations were categorized in molecular function, biological process and cellular component (Conesa and Götz, 2008). Furthermore, a macro functional classification was carried out with WEGO (Ye et al., 2018).

For assembly employing reference genomes, information reported in GDR database for *Fragaria vesca* v4 (Edger et al., 2018); *Rubus occidentalis* v1 (VanBuren et al., 2016) and v3 (VanBuren et al., 2018); and *Rosa chinensis* v1 (Saint-Oyant et al., 2018) were taken. Sequence alignment was completed including collected information, obtaining an index according to each reference genome employed with Bowtie 2 v.2.3.2 (Langmean and Salzberg, 2012), then reads were mapped into the corresponding reference genome with TopHat v2.1.1 (Trapnell et al., 2009).

Expression analysis

The best alignment was defined as the one with the major number of mapped reads and with greatest proportion of multiple alignments. From the chosen assembly, differential expression values were determined through workflow with cufflinks, cuffmerge, and cuffdiff v2.2 (Trapnell et al., 2010), the blind method was employed taking into account that only one replicate for each treatment existed, which treats every sample as replicates with a single condition. This could result in a mid-demanding method, but it shall be considered that the method will identify very few genes as differentially expressed. From this method, normalized FPKM (fragments per kb per million) values were obtained for each condition, fact that helps to avoid effects induced by different gene sizes and sequence discrepancies for gene expression assessment, allowing the comparison among different groups. Significance values shown were evaluated taking into account the paired comparison of the three groups: CSI *vs*. CTI; CSI *vs*. CSNI; and CTI *vs*. CSNI. Differentially expressed genes (DEG) were considered when p-value was smaller than the false discovering rate (FDR) after correction by multiple Benjamini-Hochberg tests. From FPKMs measures, further analysis was completed using EdgeR package of R software (R Core Team, 2018) where a descriptive analysis through a multidimensional scaling (MDS) plot was conducted aiming to identify differential expression profiles among groups. A fold-change analysis was also considered to identify global differential expression and represent them as MA (log ratio Mean Average) graphics highlighting markers with a FDR value < 0.005.

Subsequently, from significant DEGs the following strategy was proposed to define candidate genes; first, DEG between CSI and CTI was considered which could be DEGs associated to *C. gloeosporioides*-tolerance; DEGs between CSI and CSNI, which could be attributed to the very inoculation to susceptible plant material; and last, DEG between CTI *vs*. CSNI could correspond to activated genes in regard to the very inoculation to tolerant plant material. Finally, common genes differentially expressed between tolerant and susceptible material (in both conditions) were proposed as candidate genes. An ontology analysis was performed based on selected DEG (Gene Ontology Consortium).

Variant analysis

From assembly information, SAM (Sequence Alignment/Map) index/files were converted to BAM (Binary Alignment/Map), ordered and analyzed through different SAMtools v1.9 (Li, 2011) tools, filtered with –D80 to identify SNPs with read depths greater than 80. These analyses were also conducted for each treatment in order to describe the sample plant material in a better way. MISA software (Beier et al., 2017) was employed to identify SSR markers.

RESULTS AND DISCUSSION

De novo assembly

Once the inoculation of susceptible and tolerant *R. glaucus* plant material and RNA extraction stages were completed, paired end sequences were obtained. An initial analysis that comprised all sequences from the three sample groups accounted 167 962 580 sequences, with an average length of 74.4 bp and a guanine-cytosine (GC) content of 49.3%. In regard to sequence quality, adapter presence was discarded and more than 97% of reads showed a quality score greater than 20. Further statistics classified by treatment are shown in Table 1. Sequence data are available in SRA database, BioProject accession (PRJNA527868), it is important to notice that sequence quality reported in this research are adequate and similar to those transcriptome assemblies of other Rosaceae family species previously reported (Garcia-Seco et al., 2015; Jo et al., 2015; Koning-Boucoiran et al., 2015; Han et al., 2017).

After processing, *de novo* assembly of all sequences of the three groups together yielded 55 185 contigs, out of which 38 653 corresponded to the longest isoforms. The transcriptome assembly possessed a GC percentage of 44.65% with an average contig size of 443.95 bp, a N50 of 489 bp and an amount of 24 499 604 of assembled bases along the transcriptome (Table 1), the N50 is defined as the sequence length of the shortest contig at 50% of the total transcriptome length. These values differ in some extent to those reported by Garcia-Seco et al. (2015) in *Rubus* sp. Fruits, where it is observed that consensus transcript numbers are similar but with higher contig N50 values, due to the fact that different plant tissues can exhibit different expression profiles. In general, more similar results were obtained in sour cherry transcriptome assembly with 61 043 transcripts, a GC percentage of 42.89%, a N50 values of 611 bp and an average contig length of 506.6 bp (Jo et al., 2015).

De novo assembly statistics classified by treatment are shown on Table 1, where it is noticeable that when an independent assembly was tried for each group, worse results were obtained. A higher amount of contigs with different size, represents the situation fitting within the Trinity tutorial, where a RNA sequences concatenation is suggested in order to complete *de novo* assembly (Grabherr et al., 2011).

Smallest and largest sizes of contigs corresponded to 201 and 15 493 bp, with a N10 of 1816 and a marked tendency to frequency reduction when contig size increases. In regard to the analysis of possible genes a N50 value of 406 bp, an average contig size of 396.58 bp, and an amount of 15 329 062 assembled bases were obtained. From clustering analysis, 43 579 unigenes were identified.

Some authors have reported *de novo* assemblies with similar qualities to those reported when reference genomes are employed (Grabherr et al., 2011; Garcia-Seco et al., 2015; Jo et al., 2015), even Trinity has shown excellent performance in expression analysis (Grabherr et al., 2011). Nonetheless, quality results depend on genome quality, annotation and its version. In addition, genetic closeness of the reference genome used for assembly is important, in that sense, sometimes doing *de novo* assembly could be a better choice than using reference genomes with little or insignificant relation to the studied species. As a result, *de novo* assembly has been chosen as a good strategy for differential expression analysis and quantitative trait locus (QTL) identification.

Transcriptome annotation

In regard to genome annotation, the unigene BLASTX alignment allowed the identification of 35 824 sequences reported in the Nt database; 35 602 sequences in Nr database; 19 780 sequences in Swiss-prto; 17 425 sequences in KEGG; 23 641 sequences in GO, and 8824 sequences in COG. Prediction protein analysis showed 34 674 CDS, these values are similar to those reported for *Rubus* sp. (Garcia-Seco et al., 2015).

Table 1. Summary of *de novo* assembly of *Rubus glaucus* transcriptome.

Sample	Total raw reads	Length read raw/clean	GC content raw/clean	Total clean reads	Q20 (%)	Contigs	Mean length	N50	L50	Clusters (UniGenes)
Infected tolerant	62 000 000	66.8/90	45.0/41.5	46 022 646	96.5	121 647	317.7	300	81 487	95 311
Infected susceptible	51 516 018	80.1/90	51.0/43.9	45 843 084	97.2	58 652	464.1	527	45 521	46 364
Susceptible	54 446 562	77.6/90	52.0/43.9	46 978 736	97.7	58 636	461.3	522	45 353	46 382
Total	167 962 580	74.4/90	49.3/44.6	138 844 466	97.1	55 185	443.9	489	42 362	43 579

GC content: Content of guanine-cytosine, %Q20: quality percentage > 20, N50: length of the shortest contig at 50% of the total assembly, L50: number of contigs whose summed length is N50.

Transcriptome annotation analysis over Nr database (NCBI) revealed that the greatest portion of genes (85%) were related to Rosaceae species, especially strawberry (*Fragaria vesca*; 67.6%); followed by pear (*Prunus persica*; 17.4%) and another plants including grapes (*Vitis vinifera*; 2.5%), *Medicago truncatula* (1.1%), black poplar (*Populus trichocarpa*; 0.9%); and 10.5% of other species. From the evaluated sequences, 60.3% showed a similarity greater than 80%, and only 16.5% showed similarity values lower than 40%. It is important to consider that strawberry and pear are the most studied species within Rosaceae family; this explains why the greatest amount of annotated sequences corresponded to them. This, in turn, does not mean that similarity with closer species has not occurred, they are just less studied. Results shown in this paper are similar to those reported for other transcriptome analysis from different Rosaceae species. In rose, it is reported that 60.7% of sequences were related to strawberry (Koning-Boucoiran et al., 2015) and in *R. occidentalis* the 73.56% shown similarity with the same species, however, despite the high similarity rates with *F. vesca*, the assembly result with reference genome did not have quality and showed low mapping rates (7%).

Gene ontology (GO) transcriptome annotation by functional classification is shown in Figure 1, where 42.7% of unigenes were associated to cellular component category, 42.5% with molecular function, and 42.6% with biological process, with an amount of 23 641 genes reported in GO. It is important to notice that there was no expression in categories like *virion*, *protein lag*, among others. Similar results have been previously reported (Garcia-Seco et al., 2015; Koning-Boucoiran et al., 2015).

SSR analysis of sequences allowed to identify 3570 SSR markers in 2978 unigene sequences, 477 sequences possessed more than one SSR. The frequency of the repetition type is demonstrated in Figure 2. Reported values agree with results found in transcriptome analysis of *Rubus* sp. (Garcia-Seco et al., 2015). These markers could be employed in selection of tolerant plant material, evolutionary studies, hybridization o selective breeding.

SAMtool analysis of unigene sequences allowed the identification of 38 791 SNPs within the transcriptome, a little bit less than the reported by Garcia-Seco et al. (2015) in *Rubus* spp. (67 521 and 67 845), effect that suggest less diversity in *R. glaucus* plant material employed in the present study. This explanation also satisfies previous reports made in the region, which argues low variability in blackberry cultivars from Colombia Coffee Producing Region (Marulanda et al., 2012). However, out of the identified markers, 30 759 SNPs were heterozygous and 8032 were fixed. In regard to

Figure 1. Histogram of the Gene Ontology (GO) classification of *Rubus glaucus* transcriptome by GO categories: molecular function, cellular component, and biological process. Unigenes with length greater than 200 pb.

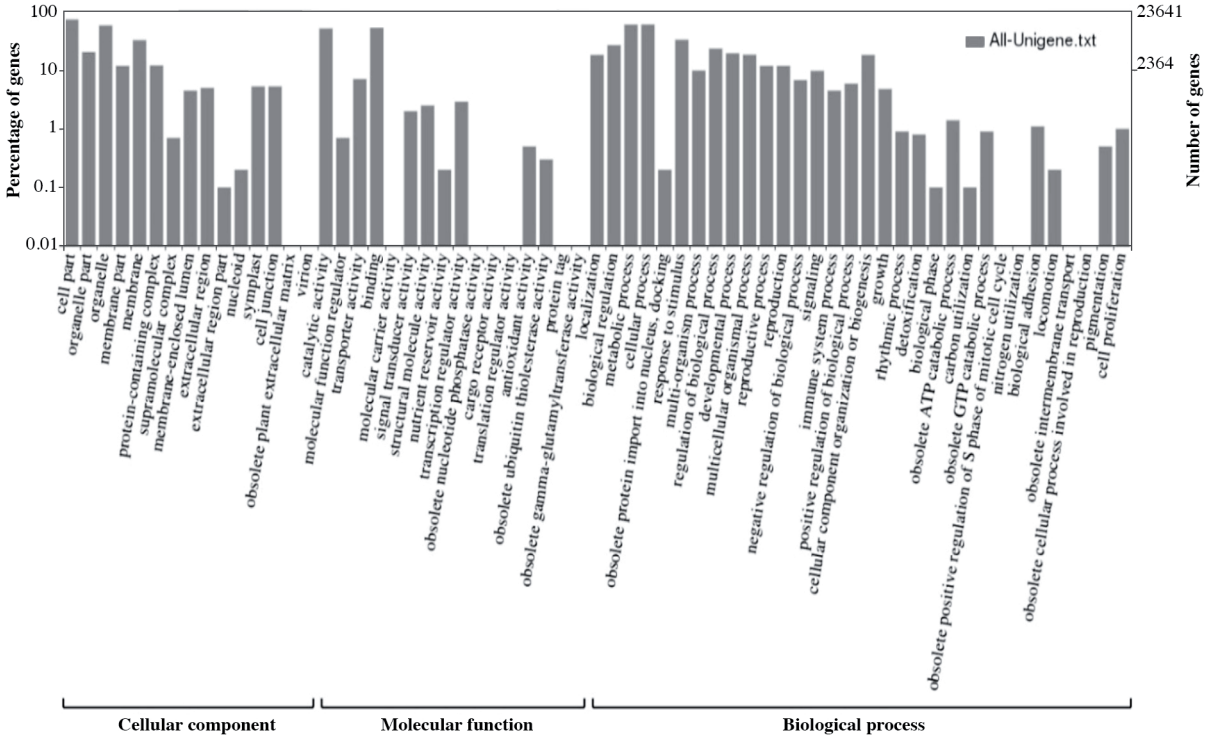
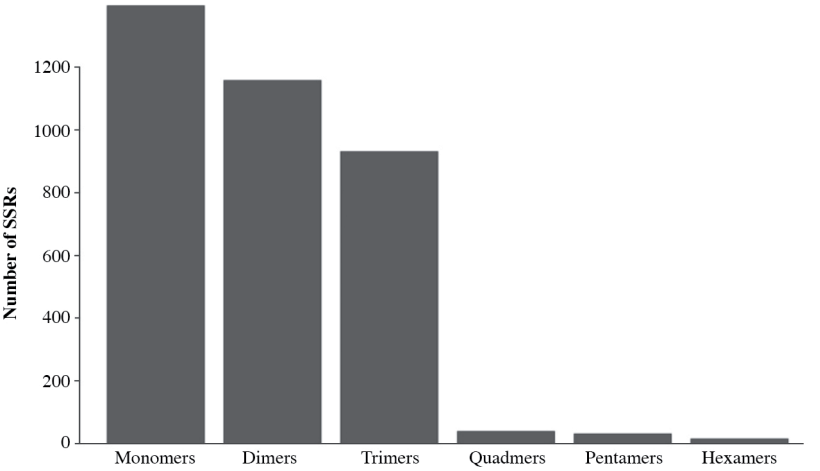


Figure 2. Frequency of repetition-types in simple sequence repeat (SSR) markers identified in the *Rubus glaucus* transcriptome.



insertions and deletions (INDELs), 1193 with more than 2 bp were observed in the transcriptome. A comparison between markers and different treatments is listed in Table 2, where a noticeable difference in the number of SNPs found in the tolerant plant material with respect the other sample groups. Also, less INDELs and heterozygous SNPs were found and, in general, less than 31.2% of common SNPs were found between the tolerant and susceptible material, which suggest genetics differences between the materials.

Assembly employing reference genomes

When a reference genome was employed, we found that the better scenario was with *R. occidentalis* v3, given that its mapping rate was around 75.5%, out of which 89.3% showed multiple alignment and an 83% concordant paired alignment rate. The worst case corresponded to the assembly using *R. chinensis* as reference genome, with a mapping of only 11.2% of the reads out of which only 13% showed multiple alignment and 2.3% were concordant paired alignments. Other assembly scenarios are shown in Table 3, were it is also clear that the last version of the assembly with *R. occidentalis* improves the number of mapped reads. Finally, in order to compare, some *de novo* statistics mappings (72.6% mapped reads, 41.8% multiple alignment and 62.4% concordant paired alignment) were also included (Table 3). From the individual mapping of each treatment, it was possible to establish that, in a general trend, CSI and CSNI showed the highest amount of mapped reads. Further statistics per treatment are shown on Table 3, where it is also shown that assembly results obtained with *F. vesca* as reference genome are not appropriate, despite the high similarity observed in annotation analysis. Garcia-Seco et al. (2015) also reported low mapping rates (7%).

The number of loci in the assembly with *R. occidentalis* as reference genome accounted 33 971, a similar result to the number of contigs in *de novo* assembly, when only large isoforms are considered (38 653), or to the number of aligned sequences in the BLASTX alignment using Nt database (35 824). Similar transcript values are reported in close species (Koning-Boucoiran et al., 2015; Han et al., 2017), suggesting a correspondence between *de novo* analysis and -with reference genome analysis.

Commonly, two different strategies for sequence analysis have been proposed, with and without reference genome. This work demonstrated that both scenarios could be robust and complementary, but when a good reference genome

Table 2. Number of single nucleotide polymorphisms (SNPs) identified in *Rubus glaucus* transcriptome according to the collection material and number of markers in common among the materials.

Sample	Total SNPs	Heterozygous	INDELs (> 2 pb)	Comparison (common SNPs)			
				1	2	3	4
1. Infected tolerant	26 064	10 113	809	-	-	-	-
2. Infected susceptible	34 491	26 403	1179	7008	-	-	-
3. Susceptible	32 275	19 196	1002	8136	11 435	-	-
4. Total	38 791	30 759	1193	14 639	19 741	17 422	-

INDELs: Insertions and deletions.

Table 3. Summary of the mapping using different reference genomes of the Rosaceae family and *de novo* assembly.

Reference genome	Sample	Read mapping rate	Reads with multiple alignments	Concordant pair alignment rate
Rose (*Rosa chinensis*)	Infected tolerant	5.5	15.8	2.3
	Infected susceptible	16.8	12.9	9.7
	Susceptible	11.3	13.7	4.7
	Total	11.2	13.0	5.6
Strawberry (*Fragaria vesca*)	Infected tolerant	20.5	15.1	14.4
	Infected susceptible	36.1	9.4	23.0
	Susceptible	39.1	10.3	24.6
	Total	32.0	10.8	20.8
Blackberry v1 (*Rubus occidentalis*)	Infected tolerant	55.6	28.6	47.5
	Infected susceptible	69.7	30.3	61.3
	Susceptible	76.4	31.9	67.0
	Total	66.9	31.1	57.6
Blackberry v3 (*R. occidentalis*)	Infected tolerant	59.5	72.9	52.1
	Infected susceptible	79.3	95.2	75.5
	Susceptible	87.2	94.9	83.0
	Total	75.5	89.3	70.5
Transcriptome (*de novo*)	Infected tolerant	48.9	42.1	40.7
	Infected susceptible	84.4	43.8	73.7
	Susceptible	84.1	40.7	72.7
	Total	72.6	41.8	62.4

is available, it could be the better alternative. On the contrary, *de novo* assembly could result in a better choice, given that in this work it demonstrated good performance and agrees with that reported by Garcia-Seco et al. (2015), with the difference that they employed *F. vesca* as reference genome. In addition, authors expressed that transcriptome of no model polyploidy species could be assembled and annotated avoiding high levels of artificial chimeras.

The analysis obtained in this work demonstrated that *de novo* analysis and –with reference genome analysis can be complementary and should be performed in order to take the best decisions if differential expression analysis is sought. This article shows that version 3 of the *R. occidentalis* has a good performance for differential expression analysis, important fact to be considered in further research. Additionally, it can be appreciated that the version 1 of the same genome (VanBuren et al., 2016) did not exhibited the same behavior, which shows the importance of further works that allow mapping and annotating different reference genomes precisely.

Differential expression

Considering that *R. occidentalis* v3 could be better annotated and aiming to improve accuracy and comparison, a fragment counting for differential expression using reference genome was developed. Mapping information helped to identify 33 971 loci for differential expression analysis, from which normalized FPKM were calculated and further quantification of differential expression was plotted onto a MDS plot which allowed to observe differences between the three sample treatments (Figure 3A). The greatest difference existed between CTI and both susceptible materials (CSI and CSNI). Some differences were observed between susceptible plant materials (inoculated and non-inoculated), suggesting that the inoculation process occurred satisfactorily and a different response in material tolerant to the fungal agent does exist (Figure 3B). Such behavior, susceptible material closeness evidencing differential response in tolerant material, suggests that involved genes could explain tolerance-traits.

The differential expression analysis, allowed the identification of 33 DEG between CTI and CSI, 5 genes between CSI and CSNI, and 22 genes between CTI and CSNI. It is important to add that these values are less that depicted in Mean Average (MA) plot because the differential expression analysis was more rigorous and not only the behavior was pursued, in that sense, genes with minor p-values than the FDR after Benjamini-Hochberg correction were selected. Out of the 33 DEG between CTI and CSI, 27 genes were up-regulated in the tolerant variety. In regard to the difference between CTI and CSNI, 18 genes were up-regulated, which means that there was up-regulation of some genes in the tolerant *R. glaucus*

variety with respect to other two samples and in almost all cases it involved up- and down-regulation on one condition, and the absence of expression in the other.

DEG analysis identified commonly expressed genes between tolerant and both susceptible varieties (CSI and CSNI) and 5 common genes that could be candidates to explain tolerance were found. Besides, from the selected genes, *RPM1* gene is reported by its importance in pathogen-plant interaction and its association with innate immune response. *RPM1* was previously identified in differential expression assays before infection (Socquet-Juglard et al., 2013). The kinase activity-related gene *MAPKBP1* possess strong importance in signal transduction and has been reported into the stress and innate immune response in plants (Suarez et al., 2010). Such gene was observed to be down-regulated in tolerant material and has also been previously reported as DEG in *Colletotrichum* infection response (Casado-Díaz et al., 2006).

On the other side, it was aimed to identify if there was some common gen between CTI and CSNI with differential expression, given that this gene could be a plausible candidate regularly expressed in the tolerant material (looked like there was no infection). The *FRL4A* gene was found to be up-regulated in CTI and CSNI, and down-regulated in CSI; this gene possesses functions related to cellular differentiation and development, and has been previously related to flowering time and multivariate adaptation (Lovell et al., 2013).

Almost all genes identified were up-regulated in tolerant material and some of them were related to cellular division and development (Table 4), fact that could be tuned with tolerance behavior of the plant material used in this work. Nonetheless, some of these genes have been little studied in plants, which raises future perspectives to be studied in-depth, as it is the case of *CDCA7* gene associated to immune response in mammals and catalogued as oncogene related to cancer process in humans (Gill et al., 2013). This kind of information could be important if it is considered that drugs used against cancer in humans could potentially exhibit activity and could be helpful to tackle plant diseases caused by fungal agents (Hadwiger and Tanaka, 2018).

Figure 3. Summary of the gene expression of *Rubus glaucus* in tolerant material inoculated with *Colletotrichum gloeosporoides* (CTI), susceptible material inoculated (CSI), and susceptible material without inoculation (CSNI). A) Multi-dimensional scaling (MDS) plot for the three conditions. B) MA (log ratio Mean Average) plot comparing the expression of CSI with CTI. C) MA plot comparing the expression of CSI with CSNI. D) MA plot comparing the expression of CTI with CSNI. In red, sequences differentially expressed with false discovering rate (FDR) < 0.005 are highlighted.

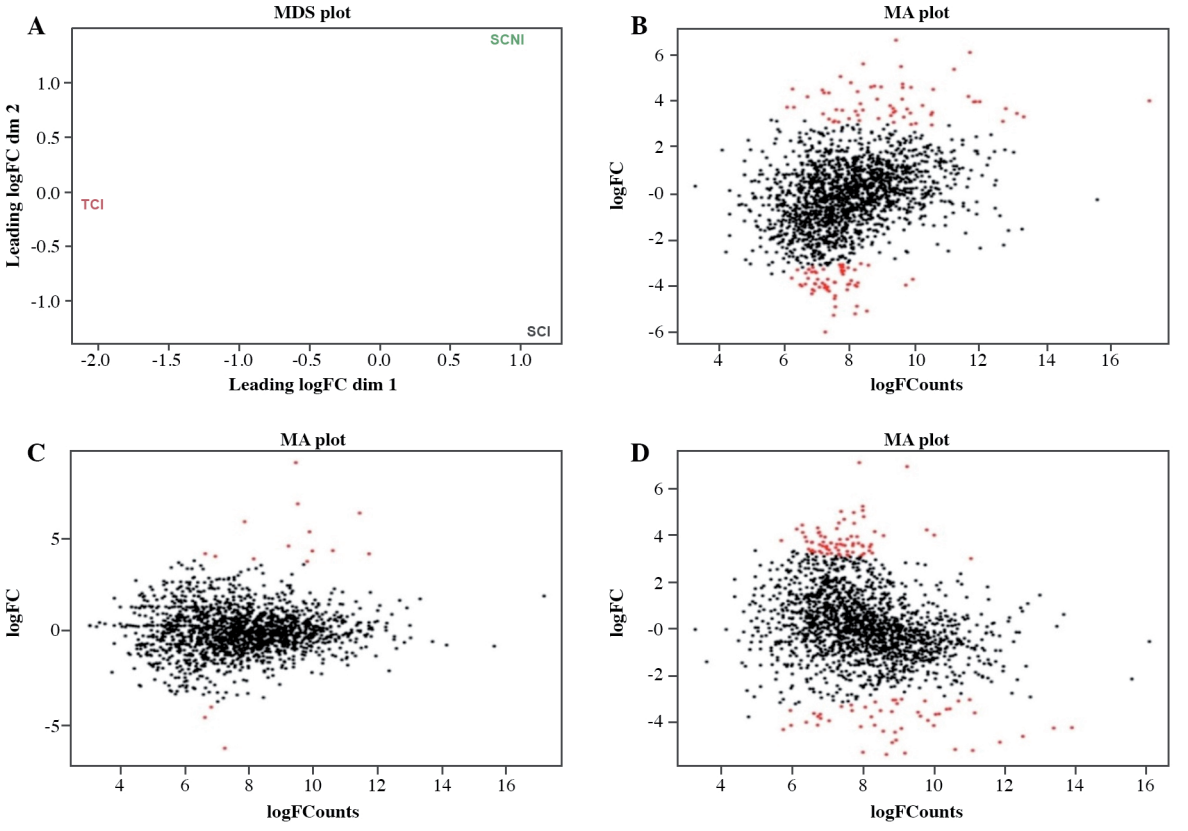


Table 4. Differentially expressed genes in material of *Rubus glaucus* tolerant to *Colletotrichum gloesporoides* with functional annotation and references of previous reports.

Loci ID (strand)	Gene name	GO description	Metabolic pathway or function	Regulated tolerant sample	Reference reported
Ro01_G02903 (+)	Cytokinin dehydrogenase 2 (*CKX2*)	Cytokinin dehydrogenase activity, metabolic process, oxidation-reduction process, catalytic activity, Flavin adenine dinucleotide binding	Zeatin biosynthesis, growth and development process in plants	Up	Associated with tiller growth and yield in rice (Yeh et al., 2015)
Ro01_G10814 (+)	B3 domain-containing protein Os1g0234100	Protein domain specific binding, DNA binding	DNA binding, regulation of transcription, transcription factor	Up	Protein family associated with processes essential for development in plants (King et al., 2013)
Ro01_G05013 (-)	Mitogen-activated protein kinase binding (*MAPKBP1*)	Protein phosphorylation, ATP binding, MAP kinase activity, protein kinase activity	Mitogen-activated protein kinase 6, Plant hormone, signal transduction and Drug metabolism – other enzymes	Down	Associated with wide array of responses, including stress, innate immunity (Suarez et al., 2010) and response to *Colletotrichum* (Casado-Díaz et al., 2006)
Ro01_G07725 (-)	Disease resistance protein (*RPM1*)	ADP binding	Plant-pathogen interaction	Up	Associated with response to the pathogen (Socquet-Juglard et al., 2013)
Ro01_G23893 (-)	Cell division cycle-associated protein 7 (*CDCA7*)	Regulation of cell proliferation, apoptosis process, regulation of transcription	Participates in MYC-mediated cell transformation and apoptosis	Up	Associated with immunodeficiency and cancer response in humans (Gill et al., 2013). No reports on plants
Ro01_G07699 (+)	FRIGIDA like-protein 4A (*FRL4A*)	Cell differentiation, growth and developmental stages	Cell differentiation	Up	Associated with flowering time and multivariate adaptation (Lovell et al., 2013)

GO: Gene ontology; MAP: mitogen-activated protein; MYC: myelocytomatosis proto-oncogene.

An important fact to consider is that in the analysis strategy employed, only common DEG in the three experimental conditions were considered, aiming to boost robustness of the analysis taking into account that biological replicates were not considered and that RNA-seq analysis could result certainly unreliable. Nevertheless, this work represent a start-point to explore confirmatory differential expression qRT-PCR analysis, with biological replicates, that make possible to identify markers for tolerant material and allow the completion of research looking for causative mutations and to perform molecular marker-assisted selection in the future.

Finally, there are different transcriptome assembly reports and differential expression analysis over species belonging Rosaceae family (Mousavi et al., 2014; Garcia-Seco et al., 2015; Jo et al., 2015; Han et al., 2017). This research presents an additional contribution to the understanding and further expression comparison between Rosaceae species; and to understand if mechanisms associated to features with economic importance, such as fungal tolerance, obey to the same expression patterns between members of the same family.

CONCLUSIONS

De novo transcriptome assembly was a good alternative for *Rubus glaucus* and such assembly could be accompanied by an analysis with a reference genome of a closely related species. In this work, it was evidenced that using the reference genome of *R. occidentalis* v3 could be very helpful for differential expression analysis in blackberry. In addition, the usage of poorly-developed versions (version 1) or references with weak relatedness could result in poorly relevant results, making *de novo* strategy analysis a better approach. It was also possible to report 43 579 transcripts in the transcriptome, which presented similarity to sequences of different Rosaceae species. Finally, it was possible to notice a differential expression profile in *R. glaucus* tolerant to *Colletotrichum gloeosporioides* and to identify some genes with possible association to the mentioned tolerance. Some of these genes have been previously reported, including RPM1 and MAPKBP1 resistance proteins. These genes shall be tested with confirmatory techniques in further studies.

ACKNOWLEDGEMENTS

Sincerely thanks to all "mora de Castilla" (Andean blackberry) growers who generously contributed to the development of the present research project and Universidad Tecnológica de Pereira for project funding.

REFERENCES

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. 2017. MISA-web: a web server for microsatellite prediction. Bioinformatics 33(16):2583-2585. doi:10.1093/bioinformatics/btx198.

Bolger, A.M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible read trimming tool for Illumina NGS data. Bioinformatics 30(15):2114-2120. doi.org/10.1093/bioinformatics/btu170.

Botero, M., Rios, G., Franco, G., Romero, M., Pérez, J., Morales, J., et al. 2002. Identificación y especialización de enfermedades asociadas a los cultivos de mora (*Rubus glaucus* Benth) en el eje cafetero. p. 87-92. In IV Seminario Nacional de Frutales de Clima Frío Moderado, Medellín. 20-22 Noviembre. Corpoica y Universidad Pontificia Bolivariana, Medellín, Colombia.

Casado-Díaz, A., Encinas-Villarejo, S., Santos, B.D., Schilirò, E., Yubero-Serrano, E.M., Amil-Ruíz, F., et al. 2006. Analysis of strawberry genes differentially expressed in response to *Colletotrichum* infection. Physiologia Plantarum 128(4):633-650. doi:10.1111/j.1399-3054.2006.00798.x.

Conesa, A., and Götz, S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics 619832. doi:10.1155/2008/619832.

Edger, P.P., VanBuren, R., Colle, M., Poorten, T.J., Wai, C.M., Niederhuth, C.E., et al. 2018. Single-molecule sequencing and optical mapping yields an improved genome of woodland strawberry (*Fragaria vesca*) with chromosome-scale contiguity. Gigascience 7(2):1-7. doi:10.1093/gigascience/gix124.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 18(23):3150-3152. doi:10.1093/bioinformatics/bts565.

Garcia-Seco, D., Zhang, Y., Gutierrez-Mañero, F.J., Martin, C., Ramos-Solano, B., Gutierrez, F., et al. 2015. RNA-Seq analysis and transcriptome assembly for blackberry (*Rubus* sp. Var. Lochness) fruit. BMC Genomics 16(1):1-11. doi:10.1186/s12864-014-1198-1.

Gill, R.M., Gabor, T.V., Couzens, A.L., and Scheid, M.P. 2013. The MYC-associated protein CDCA7 is phosphorylated by AKT to regulate MYC-dependent apoptosis and transformation. Molecular and Cellular Biology 33(3):498-513. doi:10.1128/MCB.00276-12.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29:644-652. doi:10.1038/nbt.1883.

Hadwiger, L.A., and Tanaka, K. 2018. DNA Damage and chromatin conformation changes confer nonhost resistance: A hypothesis based on effects of anti-cancer agents on plant defense responses. Frontiers in Plant Science 9:1056. doi:10.3389/fpls.2018.01056.

Han, Y., Wan, H., Cheng, T., Wang, J., Yang, W., Pan, H., et al. 2017. Comparative RNA-seq analysis of transcriptome dynamics during petal development in *Rosa chinensis*. Scientific Reports 7:43382. doi:10.1038/srep43382.

Jo, Y., Chu, H., Jin, K., Hoseong, C., Lian, S., and Won Kyong, C. 2015. *De novo* transcriptome assembly of a sour cherry cultivar, Schattenmorelle. Genomics Data 6(0):271-272. doi:10.1016/j.gdata.2015.10.013.

King, G.J., Chanson, A.H., McCallum, E.J., Ohme-Takagi, M., Byriel, K., Hill, J.M., et al. 2013. The *Arabidopsis* B3 domain protein VERNALIZATION1 (VRN1) is involved in processes essential for development, with structural and mutational studies revealing its DNA-binding surface. Journal of Biological Chemistry 293:11758-11771. doi:10.1074/jbc.M112.438572.

Koning-Boucoiran, C.F.S., Esselink, G.D., Vukosavljev, M., van 't Westende, W.P.C., Gitonga, V.W., Krens, F.A., et al. 2015. Using RNA-Seq to assemble a rose transcriptome with more than 13,000 full-length expressed genes and to develop the WagRhSNP 68k Axiom SNP array for rose (*Rosa* L.) Frontiers in Plant Science 6:249. doi:10.3389/fpls.2015.00249.

Langmean, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9(4):357-359. https://doi.org/10.1038/nmeth.1923.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987-2993. doi:10.1093/bioinformatics/btr509.

López-Vásquez, J.M., Castaño-Zapata, J., Marulanda-Ángel, M.L., y López-Gutiérrez, A.M. 2013. Caracterización de la resistencia a la antracnosis causada por *Glomerella cingulata* y productividad de cinco genotipos de mora (*Rubus glaucus* Benth.) Acta Agronómica 62(2):174-185.

Lovell, J.T., Juenger, T.E., Michaels, S.D., Lasky, J.R., Platt, A., Richards, J.H., et al. 2013. Pleiotropy of *FRIGIDA* enhances the potential for multivariate adaptation. Proceedings of the Royal Society B: Biological Sciences 280:20131043. doi:10.1098/rspb.2013.1043.

Marulanda, M., Isaza, L., y Ramirez, M. 2007. Identificación de la especie de *Colletotrichum* responsable de la antracnosis en la mora de Castilla en la región cafetera. Scientia et Technica 1(37):585-590.

Marulanda, M.L., López, A.M., Isaza, L., and López, P. 2014. Microsatellite isolation and characterization for *Colletotrichum* spp., causal agent of anthracnose in Andean blackberry. Genetic and Molecular Research 13(3):7673-7685. doi:10.4238/2014.September.26.5.

Marulanda, M., López, A.M., and Uribe, M. 2012. Molecular characterization of the Andean blackberry, *Rubus glaucus*, using SSR markers. Genetic and Molecular Research 11(1):322-331. doi:10.4238/2012.February.10.3.

Mousavi, S., Alisoltani, A., Shiran, B., Fallahi, H., Ebrahimie, E., Imani, A., et al. 2014. *De novo* transcriptome assembly and comparative analysis of differentially expressed genes in *Prunus dulcis* Mill. in response to freezing stress. PLOS ONE 9(8):1-13. doi:10.1371/journal.pone.0104541.

R Core Team. 2018. R: A language and environment for statistical computing. Available at https://www.R-project.org/. R Foundation for Statistical Computing, Vienna, Austria.

Saint-Oyant, L.H., Ruttink, T., Hamama, L., Kirov, I., Lakhwani, D., Zhou, N.N., et al. 2018. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. Nature Plants 4:473-484. doi:10.1038/s41477-018-0166-1.

Socquet-Juglard, D., Kamber, T., Pothier, J.F., Christen, D., Gessler, C., Duffy, B., et al. 2013. Comparative RNA-Seq analysis of early-infected peach leaves by the Invasive phytopathogen *Xanthomonas arboricola* pv. *pruni*. PLOS ONE 8(1):e54196. doi:10.1371/journal.pone.0054196.

Suarez, M.C., Petersen, M., and Mundy, J. 2010. Mitogen-activated protein kinase signaling in plants. Annual Review of Plant Biology 61:621-649. doi:10.1146/annurev-arplant-042809-112252.

Trapnell, C., Pachter, L., and Salzberg, S.L. 2009. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105-1111. doi:10.1093/bioinformatics/btp120.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28(5):511-515. doi:10.1038/nbt.1621.

VanBuren, R., Bryant, D., Bushakra, J.M., Vining, K.J., Edger, P.P., Rowley, E.R., et al. 2016. The genome of black raspberry (*Rubus occidentalis*). The Plant Journal 87:535-547. doi:10.1111/tpj.13215.

VanBuren, R., Wai, C.M., Colle, M., Wang, J., Sullivan, S., Bushakra, J.M., et al. 2018. A near complete, chromosome-scale assembly of the black raspberry (*Rubus occidentalis*) genome. Gigascience 7(8):1-9. doi:10.1093/gigascience/giy094.

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., et al. 2018. WEGO 2.0: A web tool for analyzing and plotting GO annotations, 2018 update. Nucleic Acids Research 46(W1):W71-W75. doi:10.1093/nar/gky400.

Yeh, S.Y., Chen, H.W., Ng, C.Y., Lin, C.Y., Tseng, T.H., Li, W.H., et al. 2015. Down-regulation of cytokinin oxidase 2 expression increases tiller number and improves rice yield. Rice 8:36. doi:10.1186/s12284-015-0070-5.

Zheng, D., and Hrazdina, G. 2010. Cloning and characterization of an expansin gene, *RiEXP1*, and a 1-aminocyclopropane-1-carboxylic acid synthase gene, *RiACS1* in ripening fruit of raspberry (*Rubus idaeus* L.) Plant Science 179(1-2):133-139. doi:10.1016/j.plantsci.2010.04.001.