Characterizing glycosyltransferases by a combination of sequencing platforms applied to the leaf tissues of Stevia rebaudian

Shaoshan Zhang
Sichuan Agricultural University
Qiong Liu
Sichuan Agricultural University
Chengcheng Lyu
Sichuan Agricultural University
Jinsong Chen
Sichuan Agricultural University
Renfeng Xiao
Sichuan Agricultural University
Jingtian Chen
Sichuan Agricultural University
Yunshu Yang
Sichuan Agricultural University
Huihui Zhang
Sichuan Agricultural University
Kai Hou
Sichuan Agricultural University
Wei Wu (ewuwei@sicau.edu.cn)
sichuan agriculture university https://orcid.org/0000-0003-3559-1974

Research article

Keywords: Stevia rebaudiana (Bertoni), next-generation sequencing, single-molecule real-time sequencing, glycosyltransferase, phylogenetic tree

Posted Date: March 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-17892/v1

Abstract

Background: *Stevia rebaudiana* (Bertoni) is considered one of the most valuable plants because of the steviol glycosides (SGs) that can be extracted from its leaves. Glycosyltransferases (GTs), which can transfer sugar moieties from activated sugar donors onto saccharide and nonsaccharide acceptors, are widely distributed in the genome of *S. rebaudiana* and play important roles in the synthesis of steviol glycosides.

Results: Six stevia genotypes with significantly different concentrations of SGs were obtained by induction through various mutagenic methods, and the contents of seven glycosides (stevioboside, Reb B, ST, Reb A, Reb F, Reb D and Reb M) in their leaves were considerably different. Then, NGS and single-molecule real-time (SMRT) sequencing were combined to analyse leaf tissue from these six different genotypes to generate a more complete and correct full-length transcriptome of *S. rebaudiana*. Two phylogenetic trees of glycosyltransferases (SrUGTs) were constructed by the neighbour-joining method and successfully predicted the functions of SrUGTs involved in SG biosynthesis. With further insight into glycosyltransferases (SrUGTs) involved in SG biosynthesis, the weighted gene co-expression network analysis (WGCNA) method was used to characterize the relationships between SrUGTs and SGs, and forty-four potential SrUGTs were finally obtained. Of these potential SrUGTs, twenty-seven have complete ORFs and have not been verified to date, and enzyme assays were subsequently performed, but none of them had activity towards SGs. In addition, SrUGT (SrUGT88B1-1) could utilize UDP-glucose as a sugar donor to glycosylate isoquercetin to form three products containing one, two, and three glucoses, respectively.

Conclusion: Combined with the results obtained by previous studies and those of this work, we systematically characterized glycosyltransferases in *S. rebaudiana* and confirmed that four enzymes (SrUGT85C2, SrUGT74G1, SrUGT76G1 and SrUGT91D2) are primarily involved in the glycosylation of steviol glucosides. Moreover, the complete and correct full-length transcriptome obtained in this study will provide valuable support for further research investigating *S. rebaudiana*.

Introduction

Stevia rebaudiana (Bertoni) belongs to the Asteraceae family and is also one of the only two members (S. rebaudiana and S. phlebophylla) in this genus to produce steviol glycosides (SGs), compounds that appeal to people looking for more natural plant-based ingredients in their diet because of their natural-origin, plant-based and zero-calorie sustainable sweeteners [1, 2]. Indeed, finding non-nutritive sweeteners to reduce sugar intake is one valuable path to solve some serious health problems at present, especially for those with obesity or diabetes. Therefore, making stevia a leaf crop with significant economic value and the food industry has a very positive outlook regarding the opportunities for SGs. Consequently, the value of the SG market is expected to exceed $1 billion USD by 2021 [3]. To date, more than thirty-five SGs have been isolated and identified from S. rebaudiana, including steviolbioside, rebaudioside A-Q (Reb A-Q), 1,2-stevioside (ST), dulcoside A, dulcoside B and rubusoside [4, 5]. Notable progress has been made

in elucidating the biosynthetic pathway of SGs [6, 7]. Taking steviol, the precursor of SGs, as an example, its biosynthesis in stevia was largely determined to include nine enzyme-catalysed reactions from isopentenyl diphosphate/dimethylallyl diphosphate [8]. With the discovery of new glycosides and the characteristics of genetic heterozygosity of stevia [9, 10], elucidating the biosynthetic pathways (especially for the enzymes in the UDPG-dependent glucosyltransferase (UGT) family, which play a critical role in the production of SGs) and regulatory mechanisms of active SGs has attracted the attention of scientists.

Owing to the interest in the properties of SGs, there has been extensive transcriptome research investigating stevia. An early report used a mean of expressed sequence tags (ESTs) to identify candidate UGTs involved in the glucosylation of SGs and successfully collected more than 5500 fully annotated ESTs from S. rebaudiana leaf; finally, three UGTs (SrUGT85C2, SrUGT74G1, SrUGT76G1) that were responsible for the glucosylation reactions leading from steviol to Reb A were obtained [6, 11]. In 2014, Chen et al. used next-generation sequencing (NGS)-based RNA-Seq technology (Illumina RNA-Seq) to sequence three stevia genotypes with different Reb A and ST contents. A total of 191590282 high-quality reads were generated, and 80160 assembled unigenes were obtained [12]. In this study, although many potential unigenes involved in the SG biosynthetic pathway were obtained, none of them had been functionally characterized. Subsequent RNA-Seq analysis (Illumina platform) of the stevia leaves in two different growing stages yielded twenty-three upregulated SrUGTs, but none of them had desired activity; however, three SrUGT91D2 cDNA variants that were not among the differentially expressed SrUGTs in this RNA-Seq were cloned from five individual plants and detected to catalyse the formation of the 1,6-β-D-glucosidic linkage of SGs [7, 13]. Moreover, there were still some other Illumina sequences of stevia [14], and these efforts provided abundant transcriptome data for stevia.

Because of technical limitations, the reported average lengths of the isotigs from the Illumina platform are < 500 bp and generally need assembly to obtain full-length transcripts, resulting in redundancy and distortion of the data [15]. Single-molecule real-time (SMRT) long-read sequencing technology (Pacific Biosciences of California, Inc, http://www.pacificbiosciences.com/), known as a third-generation sequencing platform, is the most reliable means of sequencing full-length cDNA molecules and is widely used in genome sequencing because of its long reads (average 4–8 kb), higher throughput, faster detection speed and fewer systematic errors caused by in vitro reverse transcription [16, 17]. Consequently, the use of SMRT sequencing could offer access to more complete transcriptome data [15, 18]. At the 2017 International Nutrition Conference, it was reported that the genome and the full-length sequence of three commercial stevia varieties were sequenced on the PACBIO platform and fully annotated, but the data were not published [3]. Therefore, providing the available full-length sequence for each RNA, especially for those corrected by NGS reads, plays a key role for researchers to understand and improve existing steviol glycoside biosynthesis pathways or discover new pathways or compounds and through traditional breeding for non-GMO improvement. In the current study, we combined NGS and SMRT sequencing approaches to sequence six stevia varieties with different accumulation levels of SGs and then generated a more complete/full-length transcriptome of S. rebaudiana. Accordingly, the transcriptome data obtained in this study provide a valuable resource for further research investigating of

stevia, especially for SG biosynthesis. Moreover, we also cloned thirty full-length SrUGT cDNAs, and then functional identification was performed. A composite phylogenetic tree containing all SrUGTs was also constructed.

Results

Content of glucosides in samples

All validation projects used for detecting steviolbioside, rebaudioside B (Reb B), 1,2-stevioside (ST), rebaudioside F (Reb F), rebaudioside A (Reb A), rebaudioside D (Reb D) and rebaudioside M (Reb M) from HPLC-UV analysis satisfied the quantitative requirements (Table 1). The contents of steviol glucosides in the leaves of all detected samples are shown in Table 2, and the HPLC chromatograms are also shown in Fig. 1. The results demonstrated that all of the analytical glucosides in the experimental genotypes were clearly varied and provided a potential basis for WGCNA co-expression network analysis to uncover the glucosyltransferases involved in the biosynthesis pathway of the corresponding glycosides. Furthermore, data obtained separately from the leaves of the seedling, adult, and budding stages of the ‘023’ genotype also revealed that the accumulation of steviol glycosides in S. rebaudiana peaked in the budding period.

Table 1

Validation method parameters for quantification of seven steviol glycosides.

Parameters	steviolbioside	Reb B	ST	Reb F	Reb A	Reb M	Reb D
LOD (μg/ml)	7.3	3.1	1.6	2.3	2.8	3.4	3.1
LOQ (μg/ml)	20.1	9.7	6.4	9.2	10.8	12.2	10.6
Calibration curve	y = 266.95x + 25.022	y = 219.28x + 1.9539	y = 222x + 2.7051	y = 170.87x-2.5236	y = 166.02x + 6.0046	y = 160.66x + 8.6147	y = 166.18x + 11.745
Mean correlation coefficient (R2)	0.9988	0.9989	0.9992	0.9991	0.9991	0.9995	0.9994
Linear range (μg/ml)	20.1–597.0	19.4-568.1	25.6-3610.4	33.7–586.0	21.7–4820.0	14.4–301.0	15.5–567.0
Accuracy (%, n = 3)	87.3	108.1	96.5	104.3	97.4	85.6	103.8
Injection precision (RSD%, n = 6)	1.97	1.94	1.53	2.03	1.72	1.58	2.34
Stability (RSD%)	1.51	1.73	1.21	1.93	1.05	1.46	2.01
System suitability (RSD%, n = 6)	1.92	1.85	1.74	2.13	1.59	1.95	1.87

Table 2
Contents of seven steviol glycosides in analytical samples.

Samples	Content(%)						
	Steviolbioside	Reb B	ST	Reb F	Reb A	Reb M	Reb D
023-L1	0.53	0.36	1.98	1.76	8.80	-	0.36
023-L2	1.26	0.21	2.01	1.24	8.42	-	0.24
023-L3	0.97	0.12	1.55	0.98	7.74	0.05	0.06
110-L1	0.89	-	10.49	0.26	0.19	-	0.07
110-L2	0.99	-	9.44	0.29	-	-	0.09
110-L3	1.11	-	9.75	0.16	-	-	0.05
B1188-L1	-	0.43	4.13	0.21	5.77	0.14	3.17
B1188-L2	-	0.44	4.05	0.28	6.42	0.13	3.06
B1188-L3	-	0.29	3.93	0.33	5.76	0.12	3.06
GX-L1	-	0.44	3.27	0.54	14.46	-	0.95
GX-L2	-	0.37	3.13	0.45	14.08	-	0.71
GX-L3	-	0.47	3.19	0.46	13.09	0.05	1.03
11-14-L1	1.32	0.69	2.42	1.09	8.30	0.23	0.46
11-14-L2	1.61	0.60	1.94	0.87	6.37	0.16	0.35
11-14-L3	1.90	0.51	1.87	0.85	6.44	0.18	0.29
GP-L1	-	0.43	3.31	0.46	12.75	-	0.94
GP-L2	-	0.52	3.38	0.48	13.08	-	1.01
GP-L3	-	0.48	3.61	0.51	13.65	-	1.06
Note: "-" means no detection or lower than LOD							

Output from combining sequencing approach to the leaves of stevia

To obtain the complete full-length transcriptome of S. rebaudiana and identify the potential genes involved in the biosynthesis pathway of steviol glycosides, both NGS (ILLUMINA) and SMRT (PACBIO) sequencing platforms were combined to sequence six different stevia genotypes. First, eighteen high-quality cDNA libraries from six different genotypes (each in triplicate) were sequenced on the Illumina

HiSeq X Ten platform, and 48234398 clean reads were generated after quality filtering (Table S1). Then, high-quality full-length cDNAs from the pooled RNA sample of the '023' genotype were sequenced on the PacBio Sequel platform, and a total of 16289315 subreads (approximately 29.1 billion bases) were obtained. After we performed the IsoSeq protocols (https://github.com/PacificBiosciences/IsoSeq_SA3nUP/wiki#datapub), which included Circular Consensus Sequences (CCS), Classify and Cluster, a total of 39879 consensus isoform sequences with considerably more accurate sequence information were obtained, and their average length was 1949 bp. To further correct the isoform sequences, all NGS clean reads were subsequently used to correct the 39879 consensus isoform sequences using LoRDEC software [19]. After removing the redundant sequences for all SMRT isoforms using CD-HIT (identity = 0.98) [20], 30859 nonredundant unigenes (containing approximately 59.8 million bases) were produced with a mean length of 1938 bases. In addition to unigenes coding for proteins and the remaining unigenes, homologous searches against the Coding Potential Calculator (CPC), Coding-Non-Coding Index (CNCI), Coding Potential Assessment Tool (CPAT) and Pfam reference protein databases predicted 508 of these unigenes to be long non-coding RNAs (lncRNAs) with a mean length of 1321 bases (Fig. 2A, B). After SSR analysis, the number of sequences containing SSR was 1776 and without the compound SSRs; furthermore, the repeat numbers of the mononucleotides, dinucleotides, trinucleotides, tetranucleotides, pentanucleotides, and hexanucleotides were 4212, 2030, 3544, 94, 44, and 130, respectively.

According to the results from the comparison of transcript length distribution between Illumina and PacBio Sequel platforms, it was indicated that the transcripts assembled from the Illumina short reads by Trinity software [21] could not accurately represent full-length cDNAs in S. rebaudiana. The average length of assembled unigenes from the Illumina platform (mean 905.8 bp) was notably shorter than those from the PacBio Sequel platform (mean 1949.2 bp). In addition, approximately 69.4% of the assembled unigenes from NGS reads were < 1000 bp, whereas only 13.4% of the unigenes from PACBIO reads were < 1000 bp (Fig. 3). Nevertheless, from this study, it seemed that the SMRT reads further corrected by the NGS data could be considerably better than simply relying on IsoSeq protocols. Since Iso-Seq full-length transcripts are generated directly from sequencing without assembly, these data may be used as an ideal long-read reference transcriptome of S. rebaudiana.

Phylogenetic analysis of the UDP-glycosyltransferase multigene family

UDP-glycosyltransferases (UGTs) are defined by the presence of a C-terminal consensus sequence containing 44 amino acids and are responsible for transferring a glycosyl moiety from an activated donor to an acceptor molecule in all living organisms [22, 23]. In Arabidopsis thaliana, a molecular phylogenetic tree was constructed consisting of ninety-nine UGT sequences and a composite phylogenetic tree that also includes all of the additional plant UGTs with known catalytic activities [22]. This work has significantly promoted the prediction of the evolutionary history, substrate specificities and structure-function relationships of UGTs in Arabidopsis. Nevertheless, although many studies have been performed

on the glycosyltransferases of S. rebaudiana, there are still no reports on its phylogenetic tree. Therefore, a comprehensive neighbour-joining tree with ninety-eight complete SrUGTs was constructed (Fig. 4). The alignment included twenty-six SrUGTs functionally characterized in this study. After bootstrap analysis with 1000 replicates, the SrUGTs were strongly divided into fourteen major groups, with each having a support greater than 95% in distance analysis excluding group I (66% bootstrap). The fourteen well-defined major groups of SrUGTs suggest that at one time, there were fourteen ancestral genes. One sequence (SrUGT78D2) with a long unique terminal branch, suggesting accelerated evolutionary rates, tends to distort phylogenetic analyses by reducing apparent bootstrap support for nearby clades. Therefore, the data were reanalysed without this sequence. This analysis provided stronger statistical confidence (bootstrap from 64–88%) to two of the ancestral genes, corresponding to groups M and N, which are likely to share a more recent common origin. Interestingly, a similar AtUGT78D1 gene has been found in Arabidopsis [22], indicating that the UGT78D genes in the glycosyltransferase family may have evolved more rapidly.

In an attempt to predict the structure-function relatedness of the SrUGT family, numerous UGTs identified from a wide range of plant species and having different biochemical functions were aligned with the ninety-eight SrUGTs and constructed a composite phylogenetic tree. Among these additional plant UGTs, seven of the corresponding UGTs were successfully clustered within the fourteen groups identified by this study (Fig. 5). Interestingly, PdUGT94AF1 and PdUGT94AF2 derived from Prunus dulcis involved in the formation 1,6-β-D-glucosidic linkage of Prunasin [24] were clustered in group A but had a long genetic distance between them and SrUGTs of this group, implying that the SrUGTs in this group may have no related specificity. Due to the lack of other glycosyltransferases capable of forming 1,6-glucosidic bonds, this tree cannot predict the glycosyltransferases involved in the synthesis of steviol glycosides containing 1,6-glucosidic bonds (such as Reb L). FeUGT79A8 and PhUGT79A1 identified from Fagopyrum esculentum and Petunia x hybrida could utilize UDP-rhamnose as sugar donors [25]; both UGTs belong to group A and are closely related to the SrUGT79 subfamily (100% bootstrap), demonstrating that the SrUGT79 subfamily may be the glycosyltransferases of UDP-rhamnose specificity in S. rebaudiana. In addition, EUGT11 from Oryza sativa and SrUGT91D2 are known to catalyse the 1,2-β-D-glucosidic linkage of steviol glycosides [7, 26]; therefore, it is reasonable to believe that the SrUGT91 subfamily is responsible for the formation of the 1,2-β-D-glucosidic linkage in stevia. UBGAT from Scutellaria baicalensis is one of the few glycosyltransferases identified in plants that could use UDP-GlcUA as the sugar donor to catalyse a glucuronosylation reaction [27]. In the composite tree, UBGAT was clustered in group C and had a closer relationship with the SrUGT88 subfamily; moreover, the similarity of the PSPG box between the UBGAT and SrUGT88 subfamily was more than 65%; therefore, we speculate that the SrUGT88 subfamily should be the UDP-glucuronic acid-recognizing glycosyltransferases in stevia. AtUGT78D1 identified from Arabidopsis thaliana could utilize UDP-xylose as a sugar donor [28]. In this tree, AtUGT78D1 and SrUGT78D2 are clustered in group L and have high bootstrap support (100%). Therefore, it is speculated that SrUGT78D2 should be a glycosyltransferase in stevia, which may recognize UDP-xylose and participate in the xylosylation of glycosides, such as RF. In 2005, Richman et al. (2005) identified and characterized three UGTs (SrUGT85C2, SrUGT74G1 and SrUGT76G1) involved in

the synthesis of steviol glycosides: SrUGT85C2 and SrUGT74G1 glucosylate the C13-hydroxyl and C19-carboxylic acid functional groups of the steviol backbone, forming a β-D-glucoside, respectively, while SrUGT76G1 is capable of catalysing 1,3-β-D-glucosylation at both the C13 and C19 positions of steviol. Most SrUGTs in group N belong to the SrUGT85 subfamily and have a high support value, especially for SrUGT85C3 and SrUGT85C4; therefore, we hypothesized that the SrUT85C subfamily may be responsible for glucosylating the C13-hydroxyl position of the steviol backbone. Similar to the SrUGT85 subfamily, in group J, the SrUGT74G subfamily could theoretically be the glucosylated C19-carboxylic acid functional group of the steviol backbone. Furthermore, the SrUGT76G and SrUGT76I subfamilies not only have a high degree of support (100% bootstrap) but also have a close genetic distance, indicating that these two subfamilies may be responsible for the formation of the 1,3-β-D-glucosidic linkage.

WGCNA co-expression network analysis for the investigation of steviol glycoside biosynthesis

To date, many steps involved in the biosynthesis pathway of steviol glycosides have been successfully uncovered, especially for elucidating four UDP-dependent glucosyltransferases (UGTs) [6, 29], but several glucosylation steps of some glycosides that have not been resolved to date. In addition, for the large family of glucosyltransferases, we speculate that multiple enzymes with similar functions may participate in the same catalytic step in the glucosylation of steviol glycosides. Typically, the traditional method for differential expression analysis is constrained to paired sample analysis and thus unable to perform systematic analysis with large datasets from heterogeneous sources simultaneously [30, 31]. Therefore, in this study, one co-expression network approach named WGCNA, which was proved to be a powerful tool in systematically describing the correlation relationship between clusters of highly correlated genes or modules and external conditions or sample traits [31, 32], was used to analyse the potential UGTs involved in the glucosylation of steviol glycoside. First, we performed qRT-PCR analysis of the expression levels of nine UGTs (SrUGT71H1, SrUGT85B1-2, SrUGT91D2, SrUGT76G1-1, SrUGT91D3, SrUGT85C3, SrUGT79A2, SrUGT73G2 and SrUGT71I1) in the leaves of six genotypes to confirm the reliability of the transcriptome data, and the primers used for qRT-PCR are shown in Table S2. The results showed that the tendency of these genes to be expressed was similar between the qRT-PCR and the transcriptomic data, confirming that the transcriptomic results were reliable (Fig. 6). To avoid transcript loss, first, we compared the assembled transcripts (total number 71718) from NGS with those (total number 30859) sequenced from PACBIO using BLAST + software (http://blast.ncbi.nlm.nih.gov), and the results showed that more than ninety-nine percent of PACBIO isoforms were included in the assembled transcripts of NGS (Fig. 7), indicating that the unigenes assembled by NGS contain all transcripts in S. rebaudiana. Consequently, all 71718 transcripts assembled from NGS as input raw data for WGCNA co-expression network analysis. First, genes with low fluctuation expression (standard deviation ≤ 1) were filtered, and 14995 genes remained. When the power value of adjacency functions for weighted networks was 9, both the correlation coefficient and degree of gene connectivity could satisfy the requirement of scale-free network distribution to the greatest extent possible. Based on the selected power value, a

weighted co-expression network model was established, and 14995 genes were eventually divided into fifteen modules, of which the grey module had no reference significance because of the failure to assign to any module. The hierarchical clustering dendrogram of gene networks is visualized in Fig. 8A.

To identify modules that are significantly associated with the traits of steviol glycoside content, fifteen generated modules were correlated with the traits. The modules related to each trait were screened according to the absolute value of correlation coefficient ≥ 0.3 and p-value < 0.05. The colour-coded table in Fig. 8B shows the full module-trait relationships. For each trait-related module, the correlation between the gene expression profile and the corresponding traits (Gene Significance, GS) and the correlation between the gene expression profile and the module eigengenes were calculated. The results showed that the genes in the module are both highly correlated with the traits and the eigengenes. For example, the module eigengenes of turquoise (r = 0.71, correlation p-value = 0.00092) and dark-orange (r=-0.78, correlation p-value = 0.00014) were significantly positively correlated or negatively correlated with RA, respectively. As a result, fourteen gene modules that are highly associated with steviol glycosides were identified. Among these genes in the fourteen modules, two genes belong to the acetylglucosaminyltransferase, fifty-five genes were annotated to be members of the plant UGT superfamily, including three SrUGT85C2, one SrUGT74G1, one SrUGT76G1, one SrUGT85A8 and one SrUGT91D2, which have already been reported to be involved in the glucosylation of steviol glycosides except for SrUGT85A8 [6, 7], illustrating the reliability of our results. Furthermore, the expression levels of these genes in the leaves of the six genotypes are shown in Fig. 9.

In S. rebaudiana, steviol glycosides have been derived from the tetracyclic diterpene steviol backbone [6], and the precursors of steviol are actually synthesized via a series of enzymes consisting of DXS, DXR, CMS, CMK, MCS, HDS, HDR, GGDPS, CPPS, KS, KO and KAH [8, 33, 34]. Accordingly, the genes encoding enzymes involved in steviol biosynthesis might be expected to exhibit a similar co-expression pattern with the SrUGTs involved in the synthesis of steviol glycosides. Therefore, we further performed a similar co-expression analysis between the fifty-five SrUGTs and the genes involved in steviol biosynthesis. Notably, forty-four SrUGTs, including SrUGT85C2, SrUGT74G1, SrUGT76G1 and SrUGT91D2, were then identified as being co-expressed with at least one of the upstream genes (Fig. 9), and it is reasonable to believe that these SrUGTs may be directly involved in the synthesis of the corresponding steviol glycosides, which warrants further research.

Screening candidate SrUGTs for enzyme activity

After analysis of the full-length sequence from the PACBIO platform of the forty-four candidate SrUGTs, twenty-seven SrUGTs in addition to the known SrUGTs were found with the complete ORF. In addition to one SrUGT that cloned only part of the fragment, the remaining twenty-six SrUGTs successfully isolated full-length cDNAs using specific primers (Table S3), and nucleotide and protein sequences of these SrUGTs are shown in the supplementary material. Among these twenty-six isolated SrUGTs, the recombinant enzymes from twenty-five of them and a vector control expressed successfully in E. coli, except for SrUGT73G2 (Figure S1), and the induction conditions are shown in Table S4. After purification, the bands of twenty-one SrUGTs corresponding to 55 ~ 67 kDa were clearly observed, and four proteins

(SrUGT85C3, SrUGT71H1, SrUGT88B2, SrUGT79A2) were blurred due to their poor solubility, confirming the formation of recombinant protein, whereas no such bands were observed in the control (Figure S2). The activities were screened under identical reaction conditions using UDP-glucose as the sugar donor and steviol and steviol glycosides as acceptors. The reaction mixtures were tested by HPLC, but no desired activity was detected.

Moreover, functional annotation showed that the genes in the SrUGT88B, SrUGT75E, SrUGT73E and SrUGT71E subfamilies may be involved in the glycosylation of flavonoids. A previous study showed that enzymes from SrUGT71E1-2 and SrUGT88B1-1 could glycosylate kaempferol to produce new products [6]. In this study, we also chose four SrUGTs (SrUGT88C1, SrUGT73E1-2, SrUGT75E1-2 and SrUGT88B1-1) co-expressed with upstream genes in the flavonoid biosynthesis pathway to detect their catalytic activity against rutin, apigenin, apigenin-7-o-glucoside, luteolin, luteolin-7-o-glucoside, quercetin, isoquercetin, myricetin, and epicatechin. After induction and purification, only two SrUGTs (SrUGT88B1-1 and SrUGT73E1-2) obtained soluble protein (Figure S3). Then, the reaction mixtures of these two SrUGTs were analysed by HPLC, and only one candidate, SrUGT88B1-1, was found to have activity towards isoquercetin to produce three new compounds (Fig. 10). Electrospray ionization (ESI) LC-MS in the positive ion mode was used to further characterize the three new reaction products, and the results showed three new products with a mass increase of 162 Da, 162 Da, and 162 Da (162 Da was equivalent to glucose), respectively (Figure S4). Notably, the mass of the three new products was 24 Da lower relative to glycosylated isoquercetin. We lacked authentic standards to unambiguously confirm these products.

Discussion

To reduce the intake of sucrose and other high-energy sweeteners, the demand for natural non-caloric sweeteners is increasing. Steviol glycosides from the leaves of stevia constitute such a natural alternative. To this end, several previous transcriptome studies have focused on steviol glycosides [6, 7, 12, 13, 35]; however, these studies were limited by either the number, reliability, or length of the generated sequence information, generally requiring further cloning efforts to obtain the full-length cDNA sequence for the investigation of target genes in SG biosynthesis. The taste of steviol glycosides relies on glycosylation at the C13-hydroxyl and/or C19-carboxylic acid positions of the diterpenoid steviol backbone; generally, perceived sweetness is positively correlated with the total number of glucose residues present [36]. Therefore, current research primarily focuses on the discovery and regulation of glycosyltransferases, the investigation of which largely depends on synthetic biology approaches using genes codon-optimized for recombinant expression [6, 29], which clearly demands accurate and full-length cDNAs. In this study, we combined short-read NGS and long-read SMRT sequencing of six different genotypes of stevia, which are the largest number of materials used for sequencing at all times, and then successfully generated a much more complete transcriptome of the stevia leaf (total of non-redundant unigenes = 39879, average length = 1949 bp). After correcting the SMRT reads using ILLUMINA reads, we finally obtained more high-quality full-length transcripts (total of nonredundant unigenes = 30859, average length = 1938 bp), reducing misassembly of genes and gene families with high sequence

identity. To the best of our knowledge, this report presents the first public data to characterize the structure of transcripts in S. rebaudiana, which will provide a genetic reference for further studies of stevia or other closely related species. Moreover, with the publication of genomic data of S. rebaudiana, our transcriptome data will provide strong support for the study of alternative splicing (AS) forms and alternative polyadenylation (APA) events of genes in stevia.

In previous studies, the number of S. rebaudiana samples used for sequencing was usually nine (three genotypes with three repetitions), which led to the restriction of analysis methods to paired sample analysis using differential expression analysis [7, 12, 13]. For the large family of glycosyltransferases, it is difficult to systematically reveal their role in SG synthesis. After more than a decade of study, we screened six genotypes of S. rebaudiana with significantly different concentrations or types of steviol glycosides from hundreds of wild-type and mutant plants. Based on this research, we can systematically analyse the transcriptome data of eighteen S. rebaudiana materials using the WGCNA method and then systematically reveal the correlation relationship between SrUGTs and stevioside biosynthesis in stevia for the first time. Adding the twenty-five SrUGTs verified in this study, more than forty-seven SrUGTs with full-length cDNA sequences had been successfully cloned and functionally verified, but only four of them (SrUGT85C2, SrUGT74G1, SrUGT76G1 and SrUGT91D2) were found to participate in the biosynthesis of steviol glycosides in stevia [6, 7]. Although new glucosyltransferases were reported to be involved in the synthesis of Reb A, a comparison indicated that these enzymes should be mutants of SrUGT76G1 [37, 38]. Therefore, it is reasonable to believe that these four enzymes are mainly involved in the glycosylation of steviol glucosides. At the same time, we also cloned and analysed the cDNA sequences of these four SrUGTs in different stevia materials and found that they all had multiple copies, and some copies had incomplete ORFs caused by base mutations or insertions (data not shown). Therefore, we believe that the synthesis and accumulation of steviol glucosides in stevia are primarily regulated by the expression level and the sequence variation of these four genes. Unfortunately, due to the lack of UDP-Rha and UDP-Xyl, this work could not reveal SrUGTs using UDP-Rha or UDP-Xyl as sugar donors, but we provided a high-quality analysis in a composite phylogenetic tree and obtained potential candidate SrUGTs with full-length sequences. As vital secondary metabolites in high plants, flavonoids have wide pharmacological activities, and it is of great significance to reveal the enzymes involved in flavonoid biosynthesis, such as glycosyltransferase. The large majority of sugars naturally attached to flavonoids are glucose residues. In accordance with this property, most of the enzymes identified transferred this sugar. In this study, it was found that SrUGT88B1-1 can use UDP-glucose as a sugar donor to glycosylate isoquercetin to form three products with one, two, and three glucoses, respectively. Combined with a previous report, which confirmed that SrUGT88B1-1 could glycosylate kaempferol [6], SrUGT88B1-1 could recognize a variety of flavonoids as acceptors and could glycosylate multiple active sites. Our study provides a template for investigating secondary metabolism in other species, paving the way for synthetic biology approaches to such natural products. Moreover, some of these six genotype stevias could be ideal materials for industrial production and could be directly used to extract high-purity steviol glycosides.

Conclusion

In summary, the more complete and correct full-length transcriptome data of S. rebaudiana were generated by a combination of Illumina and SMRT sequencing platforms. We systematically revealed the glycosyltransferase in S. rebaudiana and confirmed that four enzymes (SrUGT85C2, SrUGT74G1, SrUGT76G1 and SrUGT91D2) are mainly involved in the glycosylation of steviol glucosides using WGCNA, phylogenetic trees, qRT-PCR and enzyme assay methods. In addition, a SrUGT (SrUGT88B1-1) could utilize UDP-glucose as a sugar donor to glycosylate isoquercetin to form three products containing one, two, three glucoses, respectively, was also found. The present study may serve as a valuable resource for future S. rebaudiana studies and may also benefit investigations involving other closely related species. The full-length transcriptome dataset may also provide useful candidate genes for the elucidation of the mechanism of steviol glucosides biosynthesis.

Methods

Plant materials and RNA sample preparation

Six *S. rebaudiana* genotypes (named 110, 11-14, 023, GX, GP, B1188) with different concentrations of the steviol glycosides were harvested from an experimental field of Sichuan Agricultural University (Chengdu, Sichuan Province in China). Plants with these genotypes except B1188 were obtained by the induction of the 11-14 genotype through various mutagenic methods. Eighteen independent samples (six different genotypes of leaf tissues with three repetitions) of the 3rd leaf tissues in the budding period (the accumulation of steviol glycosides peaked) were collected. Genotype '023', which was obtained from induction by ^{60}Co γ-ray irradiation and with a wide variety of steviol glycosides, was used for the full-length transcriptome sequence. To obtain sufficient full-length transcriptome sequences, samples of the 3rd leaf tissues of the '023' genotype were separately collected from the seedling and adult stages. All samples were identified by Professor Wei Wu, who studied *stevia* for more than 10 years, and all genotypes used in this study were planted at Sichuan Agricultural University. Material collection was conducted in accordance with local legislation, and there was no need for permission from other organizations. We complied with the Convention on the Trade in Endangered Species of Wild Fauna and Flora.

A total of twenty RNA samples were isolated using the mirVana miRNA Isolation Kit (Ambion) following the protocol; these samples were from the '110' (budding period), '11-14' (budding period), '023' (budding period), 'GX' (budding period), 'GP' (budding period), 'B1188' (budding period), '023' (seedling period), and '023' (adult period). The enrichment of Poly(A) RNA from the total RNA was carried out by the oligo d(T) magnetic bead binding method. The total RNA was quantified, and the quality was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and NanoDrop (Thermo Fisher Scientific, USA). The samples with RNA integrity number (RIN) ≥ 7 were subjected to subsequent analysis. Finally, equal amounts of '023' (seedling period), '023' (adult period) and '023' (budding period) were combined to provide the total *S. rebaudiana* RNA and then subjected to Pacific Biosciences (PacBio) single-molecule long-read sequencing (Pacific Bioscience, Menlo Park, USA). Eighteen samples of the six

different genotypes were submitted for second-generation transcriptome sequencing using the Illumina HiSeq X Ten platform (Illumina, USA).

In *S. rebaudiana*, more than thirty-five steviol glycosides have been successfully identified [4, 5]. Of these glycosides, the standards of steviolbioside, Reb B, ST, Reb F, Reb A, Reb D and Reb M were easy to purchase; furthermore, their concentrations were relatively higher than those of other glycosides. Therefore, analysis of the eight steviol glycosides in all samples was conducted by reference to the HPLC-UV method in [39]. In addition, a calibration curve, limit of detection (LOD), limit of quantification (LOQ), system suitability, precision and accuracy parameters were used to validate the method. The LOD and LOQ were calculated using both the values of the calibration curve and signal-to-noise ratios of 10 and 3, respectively [40].

Illumina short-read sequencing and transcriptome assembly

The libraries were constructed using the TruSeq Stranded mRNA LTSample Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. First, fragmented mRNAs were generated by adding the interruption reagent. Second, first-strand cDNA was synthesized by SuperScript II Reverse Transcriptase using random primer 6. Next, second-strand cDNA synthesis was performed using Phusion High-Fidelity DNA Polymerase. Purified cDNA was normalized by end repair, adenylation of the 3′ ends and ligation of the adapters. Finally, the normalized cDNA was amplified using the PCR method to enrich cDNA fragments, and the amplified cDNAs were purified by the AMPure XP system (Beckman Coulter, Beverly, USA). After purification, the PCR products were validated on an Agilent 2100 Bioanalyzer (Agilent Technologies), and the sizes were also checked by agarose gel electrophoresis. Then, these libraries were sequenced on the Illumina sequencing Illumina HiSeq X Ten platform (Illumina), and 150-bp paired-end reads were generated.

Library preparation and PACBIO sequencing

The library construction and PacBio sequencing were performed according to the official protocol as described by Pacific Biosciences (Pacific Biosciences, USA). Briefly, 1 μg of total RNA was used as input for first-strand cDNA synthesis using a SMARTer PCR cDNA Synthesis kit (Clontech, USA). The first-strand products were diluted to an appropriate volume and subsequently used for large-scale PCR. Next, a total of 12 PCR cycles of amplification were performed for second-strand cDNA synthesis using PrimeSTAR GXL DNA Polymerase (Clontech, USA). After amplification, the PCR products were purified with AMPure PB Beads (Pacific Biosciences) and then normalized by repairing DNA damage, repairing ends and blunt ligation reactions. The normalized cDNA products were then subjected to the construction of SMRTbell template libraries using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). Finally, two SMRT cells were sequenced on a PacBio Sequel instrument using sequencing kit 2.1 (Pacific Biosciences) with 10 h movie recordings.

Isoform analysis

Subreads were subjected to circular consensus sequences (CCS) using SMRT analysis software (https://www.pacb.com/products-and-services/analytical-software/devnet/), and then full-length (FL) transcripts with a correction accuracy greater than 99% (high-quality isoforms) were obtained. Pac PacBio reads were classified into full-length (FL) and non-full-length reads, and then reads were corrected with the data generated with Illumina HiSeq X Ten using LoRDEC. The corrected isoforms were clustered using CD-HIT (identity=98%) to generate nonredundant unigenes.

Functional annotation of unigenes

Nonredundant unigene sequences were selected to map with the eight databases using diamond software (Crystal Impact GbR, Germany) and HMMER software (www.hmmer.org), and obtained the annotation information of the unigenes with e-values < $1e^{-5}$ against the eight databases, including non-redundant protein sequence database (NR; https://www.ncbi.nlm.nih.gov/), Clusters of Orthologous Groups of proteins/euKaryotic Ortholog Groups (COG/KOG; ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kyva), Gene Ontology (GO; http://www.geneontology.org/), non-redundant protein sequence database (Swiss-Prot; http://www.uniprot.org/), evolutionary genealogy of genes: Nonsupervised Orthologous Groups (eggNOG; http://eggnog.embl.de/), Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/pathway.html), and the database of Homologous protein family (Pfam; http://pfam.xfam.org/).

LncRNA prediction

LncRNAs are a kind of RNA molecule measuring over 200 bp and having no coding ability. In this study, nonredundant unigene sequences were used to predict the potential lncRNAs and followed these steps: first, transcripts with ORF lengths between 200 and 300 bp were screened out of the sequences already annotated on the coding libraries; finally, the coding ability of screened transcripts was predicted using CPC, CNCI, CPAT and Pfam (v1.5) protein structure domain analysis.

Simple sequence repeat (SSR) analysis

Simple sequence repeat (SSR), from DNA sliding and mismatch during DNA replication and repair or the unequal exchange of sister chromatids in mitosis and meiosis, is a tract of DNA (ranging in length from 2-13 base pairs) with certain motifs repeated 5-50 times [41]. Nonredundant transcript sequences were selected for SSR analysis using MIcroSAtellite software (MISA, v1.0). After analysis, a total of six SSR types can be detected, including mononucleotide, dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide SSRs. In addition, the repeat times were divided into 5, 6, 7, 8, 9, 10, 11, and > 11 times.

Differential expression analysis

In this study, reads from eighteen samples of six different genotypes (110, 11-14, 023, GX, GP and B1188) were produced. The expression analysis from ILLUMINA reads of different samples was carried out with bowtie2 software (obtaining the reads on unigene for each sample) and eXpress software (calculating the FPKM value of expression).

qRT-PCR analysis

To verify the reliability of the transcriptional data, qRT-PCR experiments were carried out with ten *UGT* unigenes. Eighteen RNA samples were extracted separately from six different genotypes (110, 11-14, 023, GX, GP and B1188, one genotype with three repetitions), and the leaves used for RNA isolation were the same as those for sequencing. Reverse transcription was performed with a HiScript® II Q RT SuperMix for qPCR (+gDNA wiper) Kit (Vazyme, China). qRT-PCR primers were designed from the specific domain of target genes using PRIMER PREMIER 6 (PREMIER Biosoft, Canada), and their specificity was checked by PCR. qRT-PCR amplification was carried out in triplicate using 2×T5 Fast qPCR Mix (SYBR Green I) (TSINGKE), with 26S rRNA as the reference gene, by a 7500 real-time PCR system (ABI).

Phylogenetic analysis of *UGTs*

To maximize the transcript information of *UGT* genes, we combined all the *UGT* genes obtained from Pac PacBio sequencing and the non-redundant ILLUMINA *UGT* unigenes with Pac PacBio data. Then, those containing a single nucleotide substitution, insertion, or deletion were excluded, causing interruption of ORFs. Moreover, numerous UGTs identified from a wide range of plant species with different biochemical functions were also selected from the National Center for Biotechnology Information (NCBI) and then pooled with the abovementioned UGTs before performing an alignment using MEGA-X (MEGA, http://www.megasoftware.net/) to construct another composite tree. Furthermore, the alignment was then reconciled and further optimized to minimize insertion/deletion events. An unrooted phylogenetic tree was then constructed by the neighbour-joining clustering method with the amino acid sequences of the ORFs using the bootstrap method with 1000 replicates. Guided by the system established by the GT Nomenclature Committee [42] and further combining the classification of *UGTs* in *Arabidopsis* and the named *UGTs* in *S. rebaudiana*, we systematically classified and named *UGTs* from *S. rebaudiana*.

WGCNA co-expression network analysis

Due to the highly multivariate and complex RNA-Seq data, our first application of WGCNA to stevia transcriptome datasets was performed to reveal possible transcript modules associated with the biosynthesis of steviol glycosides, especially for *UGT* genes that directly catalyse their formation. The R package WGCNA was used to complete statistical analysis. Unsigned, weighted correlation networks

were constructed by R package WGCNA with the default power of nine. In addition, network visualization was performed using R language and Python.

Enzyme assay

The ORFs of *SrUGT* genes obtained from the WGCNA co-expression network analysis were cloned into the pET28a expression vector using the designed primers shown in Table S3. The recombinant vectors harbouring *SrUGT* genes were transferred into the prokaryotic *E. coli* expression strain BL21(DE3). All methods of protein purification, protein concentration determination, activity assay, and product analysis were performed according to the methods in [39], except for the substrates in the activity assay, which were either steviol, steviolbioside, Reb B, ST, Reb F, Reb A, Reb D, Reb M and Reb C. In addition, the substrates in the activity assay for genes in the flavonoid biosynthesis pathway were either rutin, apigenin, apigenin-7-o-glucoside, luteolin, luteolin-7-o-glucoside, quercetin, isoquercetin, myricetin, and epicatechin. HPLC analysis for the product of flavonoids was performed using a Diamonsil® C18 column (4.6 × 250 mm, 5 μm particles, Phenomenex, USA) with the following conditions: flow rate, 1 mL/min; column temperature, 40°C; injection volume, 10 μL; and wavelength, 210 nm. The mobile phases comprised a two-element system composed of acetonitrile (A) and water with 0.4% phosphoric acid (B) for gradient elution, which was performed under the following conditions: 0.0-5.0 min linear gradient, 95% B; 5.0-30.0 min, 95-0% B; and 30.0-35.0 min, 100% B. The method of protein expression for each *SrUGT* gene is shown in Table S4.

Abbreviations

SG: steviol glycoside; GTs: Glycosyltransferases; UGT: UDP-glycosyltransferase; SMRT: single-molecule real-time; NGS: next-generation sequencing; SrUGTs：Glycosyltransferases of Stevia rebaudiana; HPLC: High performance liquid chromatography; WGCNA: weighted gene co-expression network analysis; ESTs: expressed sequence tags; Reb A: rebaudioside A; Reb B: rebaudioside B; Reb D: rebaudioside D; Reb F: rebaudioside F; Reb M: rebaudioside M; ST: 1,2-stevioside; CCS: Circular Consensus Sequences; CPC: Coding Potential Calculator; CNCI: Coding-Non-Coding Index; CPAT: Coding Potential Assessment Tool; lncRNAs: long non-coding RNAs; qRT-PCR: quantitative Real-time PCR; DXS: deoxyxylulose-5-phosphate synthase; DXR: deoxyxylulose-5-phosphate reductoisomerase; CMS: 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase; CMK4-diphosphocytidyl-2-C-methyl-D-erythritol kinase; MCS: 4-diphosphocytidyl-2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; HDS: 1-hydroxy-2-methyl-2(E)-butenyl 4-diphosphate synthase; HDR: 1-hydroxy-2-methyl-2(E)-butenyl 4-diphosphate reductase; GGDPS: geranylgeranyl diphosphate synthase; CPPS: copalyl pyrophosphate synthase; KS: ent-kaurene synthase; KO: kaurene oxidase; KAH: kaurenoic acid 13-hydroxylase; LC-MS: electrospray ionization mass spectrometry; AS: alternative splicing; APA: alternative polyadenylation; UDPG: Uridine diphosphate glucose; PCR: polymerase chain reaction.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors agreed to publish.

Availability of data and material

Additional supporting information may be found in the online version of this article.

Figure S1. SDS-PAGE analysis of recombinant proteins of the twenty-five candidate SrUGTs expressed in the prokaryotic system. The arrow indicates the target protein.

Figure S2. SDS-PAGE analysis of the purified recombinant of the twenty-five candidate SrUGTs. The arrow indicates the target protein. Although we optimized the induction and purification conditions, it is also difficult to purify a large amount of clean target proteins due to the small amount of soluble protein expressed by *SrUGTs* and the background protein of the expressed strain.

Figure S3. SDS-PAGE analysis of the two flavone glycosyltransferases. A. SDS-PAGE analysis of the crude enzyme, with marker, control, SrUGT88B1-1 and SrUGT73E1-2 from left to right. B. SDS-PAGE analysis of the crude enzyme, with marker, control, SrUGT88B1-1 and SrUGT73E1-2 from left to right. The arrow indicates the target protein.

Figure S4. LC-MS analysis of the molecular weight of products catalysed by SrUGT88B1-1 using electrospray ionization (ESI) LC-MS in the positive ion mode.

Table S1. Summary of transcriptome data sequenced by the Illumina platform and their pre-treatment.

Table S2. Primers and annealing length in qRT-PCR.

Table S3. Primers and annealing temperature of the thirty *SrUGT*s.

Table S4. Induction conditions of the thirty *SrUGT*s.

Supplementary information, protein sequences used for phylogenetic tree in this study.

Supplementary information, nucleotide and protein sequences cloned by RT-PCR in this study.

The raw data from the Illumina HiSeq X Ten platform have been submitted to the Sequence Read Archive (SRA) of the NCBI under accession numbers SRR10799213, SRR10799212, SRR10799203, SRR10799202, SRR10799201, SRR10799200, SRR10799199, SRR10799198, SRR10799197, SRR10799196, SRR10799211, SRR10799210, SRR10799209, SRR10799208, SRR10799207,

SRR10799206, SRR10799205 and SRR10799204. The accession number of the raw data from SMRT sequencing was SRR10567116.

Competing interests

All authors declare that they have no competing interests for this work.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC grant no. 31671757). The supporters had no role in study design, data collection, data analysis, the writing of the manuscript or decision to publish.

Author contributions

SZ performed most of the experiments and data analysis and wrote the manuscript. QL, CL, JC, RX and JC induced and collected the materials. YY and HZ helped to plant the materials. KH and WW designed and coordinated the studies. All authors have read and approved the final manuscript.

All authors work/study in Agronomy College, Sichuan Agricultural University, Chengdu 611130, China.

Acknowledgements

We thank Shanghai OE Biotech Inc. (Shanghai, China) for high-throughput sequencing service and bioinformatics support.

References

1. Kinghorn AD, and Soejarto DD. Current status of stevioside as a sweetening agent for human use[J]. Economic and medicinal plant research / edited by H. Wagner, Hiroshi Hikino, Norman R. Farnsworth. 1985;1-52.
2. Kasai R, Kaneda N, Tanaka O, Yamasaki K, Sakamoto I, Morimoto K, Okada S, Kitahata S, and Furukawa H. Sweet diterpene-glycosides of leaves of S. rebaudiana Bertoni. Synthesis and structure-sweetness relationship of rebaudiosides-A, D, E and their related glycosides[J]. *Nippon Kagakukaishi*. 1981;5: 726-735.
3. Stephen ES, Jing R, Ong SS, Wong YY, *et al*. Insights from the Sequencing and Annotation of the Stevia rebaudian Genome and their Application in Agronomy and Health[C]//International congress of Nutrition “From Science to Nutrition Security”. October 2017.
4. Ceunen S, and Geuns JM. Steviol glycosides: chemical diversity, metabolism, and function[J]. *NAT. PROD.* 2013;76:1201-28.
5. Espinoza MI, Vincken JP, Sanders M, Castro C, Stieger M, and Agosin E. Identification, Quantification, and Sensory Characterization of Steviol Glycosides from Differently Processed *rebaudiana*

Commercial Extracts[J]. *J. Agric. Food. Chem.* 2014;62:11797-11804.

6. Richman A, Swanson A, Humphrey T, Chapman R, McGarvey B, Pocs R, and Brandle J. Functional genomics uncovers three glycosyltransferases involved in the synthesis of the major sweet glycosides of S. rebaudiana[J]. *The Plant Journal*. 2005;41:56-67.
7. Wang J, Li S, Xiong Z, and Wang Y. Pathway mining-based integration of critical enzyme parts for de novo biosynthesis of steviolglycosides sweetener in Escherichia coli[J]. *Cell Research*. 2015.
8. Brandle JE, and Telmer PG. Steviol glycoside biosynthesis[J]. *Phytochemistry*, 2007;68(14):1855-1863.
9. Miyagawa H, Fujioka N, Kohda H, Yamasaki K, Taniguchi K, and Tanaka R. Studies on the tissue culture of S. rebaudiana and its components; (II). Induction of shoot primordia [J]. *PLANTA MED.* 1986;52:321-323.
10. Sharma S, Walia S, Singh B, and Kumar R. Comprehensive review on agro technologies of low-calorie natural sweetener stevia (*rebaudiana* Bertoni): a boon to diabetic patients[J]. *J. Sci. Food. Agric.* 2016;96:1867-1879.
11. Brandle J, Richman A, Swanson AK, and Chapman BP. Leaf ESTs from *rebaudiana*: a resource for gene discovery in diterpene synthesis[J]. *Plant Mol. Biol.* 2002;50:613-622.
12. Chen J, Hou K, Qin P, Liu H, Yi B, Yang W, and Wu W. RNA-Seq for gene identification and transcript profiling of three *rebaudiana* genotypes[J]. *BMC Genomics*. 2014;15(1):571.
13. Houghton-Larsen J, Hicks PM, Naesby M, Ostergaard TT, Hansen J, Dalgaard Mikkelsen M, Halkjaer HE, Simon E, and De Andrade S. RECOMBINANT PRODUCTION OF STEVIOL GLYCOSIDES. WO. 2016.
14. Singh G, Singh G, Singh P, Parmar R, Paul N, Vashist R, Swarnkar MK, Kumar A, Singh S, Singh AK, Kumar S, and Sharma RK. Molecular dissection of transcriptional reprogramming of steviol glycosides synthesis in leaf tissue during developmental phase transitions in *rebaudiana* Bert[J]. *Scientific Reports*. 2017;7(1):11835.
15. Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, Zhu Y, Ji A, Zhang B, Hu S, Au KF, Song J, and Chen SL. Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis[J]. *The Plant Journal*, 2015;82(6):951-961.
16. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, and Eichler Resolving the complexity of the human genome using single-molecule sequencing[J]. *Nature*, 2014;517:608-611.
17. Chen SY, Deng F, Jia X, Li C, and Lai SJ. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing[J]. *Rep*. 2017;7(1): 7648.
18. Sharon D, Tilgner H, Grubert F, and Snyder M. A single-molecule long-read survey of the human transcriptome[J]. *Biotechnol*. 2013;32(3):1009-1014.
19. Salmela L, and Rivals E. LoRDEC: accurate and efficient long read error correction[J]. *Bioinformatics*. 2014;30(24):3506-3514.

20. Li W, Jaroszewski L, and Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases[J]. *Bioinformatics*, 2001;17(3):282-283.
21. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data[J]. *Nature biotechnology*. 2001;29(7):644-652.
22. Li Y, Baldauf S, Lim EK, and Bowles DJ. Phylogenetic Analysis of the UDP-glycosyltransferase Multigene Family of *Arabidopsis thaliana*[J]. *Journal of Biological Chemistry*. 2001;276(6):4338-4343.
23. Lairson LL, Henrissat B, Davies GJ, and Withers SG. Glycosyltransferases: Structures, Functions, and Mechanisms[J]. Annual Review of Biochemistry. 2008;77(1):521-555.
24. Thodberg S, Cueto JD, Mazzeo R, Pavan S, Lotti C, Dicenta F, Neilson EHJ, Møller BL, and Sánchez-Pérez R. Elucidation of the Amygdalin Pathway Reveals the Metabolic Basis of Bitter and Sweet Almonds (*Prunus dulcis*)[J]. *Plant Physiol*. 2018;178:1096-1111.
25. Eiki K, Soichiro O, Yoshinori M, Hideyuki S, Makoto S, and Goro T. Identification and characterization of a rhamnosyltransferase involved in rutin biosynthesis in\r, *Fagopyrum esculentum*\r, (common buckwheat)[J]. *Bioscience, Biotechnology, and Biochemistry*. 2018;82(10):1790-1802.
26. Yang Y, Fei L, Li J, Yang J, and Wang Y. Production of rebaudioside D by recombinant escherichia coli whole cell catalyst[J]. *Industrial Microbiology*. 2017;47(5):1-7.
27. Nagashima S, Hirotani M, and Yoshikawa T. Purification and characterization of UDP-glucuronate: baicalein 7-O-glucuronosyltransferase from Scutellaria baicalensis Georgi. cell suspension cultures[J]. *Phytochemistry*. 2000;53(5):533-538.
28. Pandey RP, Malla S, Simkhada D, Kim BG, and Sohng JK. Production of 3-O-xylosyl quercetin in Escherichia coli[J]. *Applied Microbiology and Biotechnology*, 2013;97(5):1889-1901.
29. Olsson K, Carlsen S, Semmler A, Simón E, Mikkelsen MD, and Møller BL. Microbial production of next-generation stevia sweeteners[J]. *MICROB CELL FACT*. 2016;15(1):207.
30. Ruffalo M, Koyuturk M, and Sharan R. Network-based integration of disparate omic data to identify “silent players” in cancer[J]. *PLoS Computational Biology*. 2015;11(12):e1004595.
31. Pei G, Chen L, Zhang W. WGCNA Application to Proteomic and Metabolomic Data Analysis[J]. *Methods in Enzymology*. 2017;585:135-158.
32. Langfelder P, and Horvath S. WGCNA: an R package for weighted correlation network analysis[J]. *BMC Bioinformatics*. 2008;9(1):559.
33. Lichtenhalter HK. The 1-deoxy-D-xylulose-5-phosphate pathway of isoprenoid biosynthesis in plants[J]. *Rev. Plant Physiol. And Plant Mol. Biol.* 1999;50(1):47-65.
34. Totte´ N, Charon L, Rohmer M, Compernolle F, Baboeuf I, and Geuns JMC. Biosynthesis of the diterpenoid steviol, an ent-kaurene derivative from S. rebaudiana Bertoni, via the methylerythritol phosphate pathway[J]. *Tetrahedron Lett.* 2000;41(33):6407-6410.

35. Brandle J, Richman A, Swanson AK, and Chapman BP. Leaf ESTs from *rebaudiana*: a resource for gene discovery in diterpene synthesis[J]. *Plant Mol. Biol.* 2002;50:613-622.
36. Hellfritsch C, Brockhoff A, Stähler F, Meyerhof W, and Hofmann T. Human psychometric and taste receptor responses to steviol glycosides[J]. *Agric. Food. Chem.* 2012;60(27):6782-6793.
37. Madhav H, Bhasker S, and Chinnamma M. Functional and structural variation of uridine diphosphate glycosyltransferase (UGT) gene of *rebaudiana–UGTSr* involved in the synthesis of rebaudioside A[J]. *Plant Physiol Biochem.* 2013;63:245-253.
38. Yang YH, Huang SZ, Han YL, Yuan HY, Gu CS, and Zhao YH. Base substitution mutations in uridinediphosphate-dependent glycosyltransferase 76G1 gene of *rebaudiana* causes the low levels of rebaudioside A: mutations in UGT76G1, a key gene of steviol glycosides synthesis[J]. *Plant Physiology & Biochemistry*. 2014;80:220-225.
39. Zhang S, Chen H, Xiao J, Liu Q, Xiao F, and Wu W. Mutations in the uridine diphosphate glucosyltransferase 76G1 gene result in different contents of the major steviol glycosides in *rebaudiana*[J]. *Phytochemistry*. 2019;162:141-147.
40. Saran S, Menon S, Shailajan S, and Pokharna P. Validated RP-HPLC method to estimate eugenol from commercial formulations like Caturjata Churna, Lavangadi Vati, Jatiphaladi Churna, Sitopaladi Churna and clove oil[J]. *Journal of Pharmacy Research*. 2013;6(1):53-56.
41. Gulcher J. Microsatellite markers for linkage and association studies[J]. *Cold Spring Harb Protoc*. 2012;4:425-432.
42. Mackenzie PI, Owens IS, Burchell B, Bock KW, Bairoch A, and Belanger A. The UDP glycosyltransferase gene superfamily: recommended nomenclature update based on evolutionary divergence[J]. *Pharmacogenetics*. 1997;7(4):255-269.

Figures

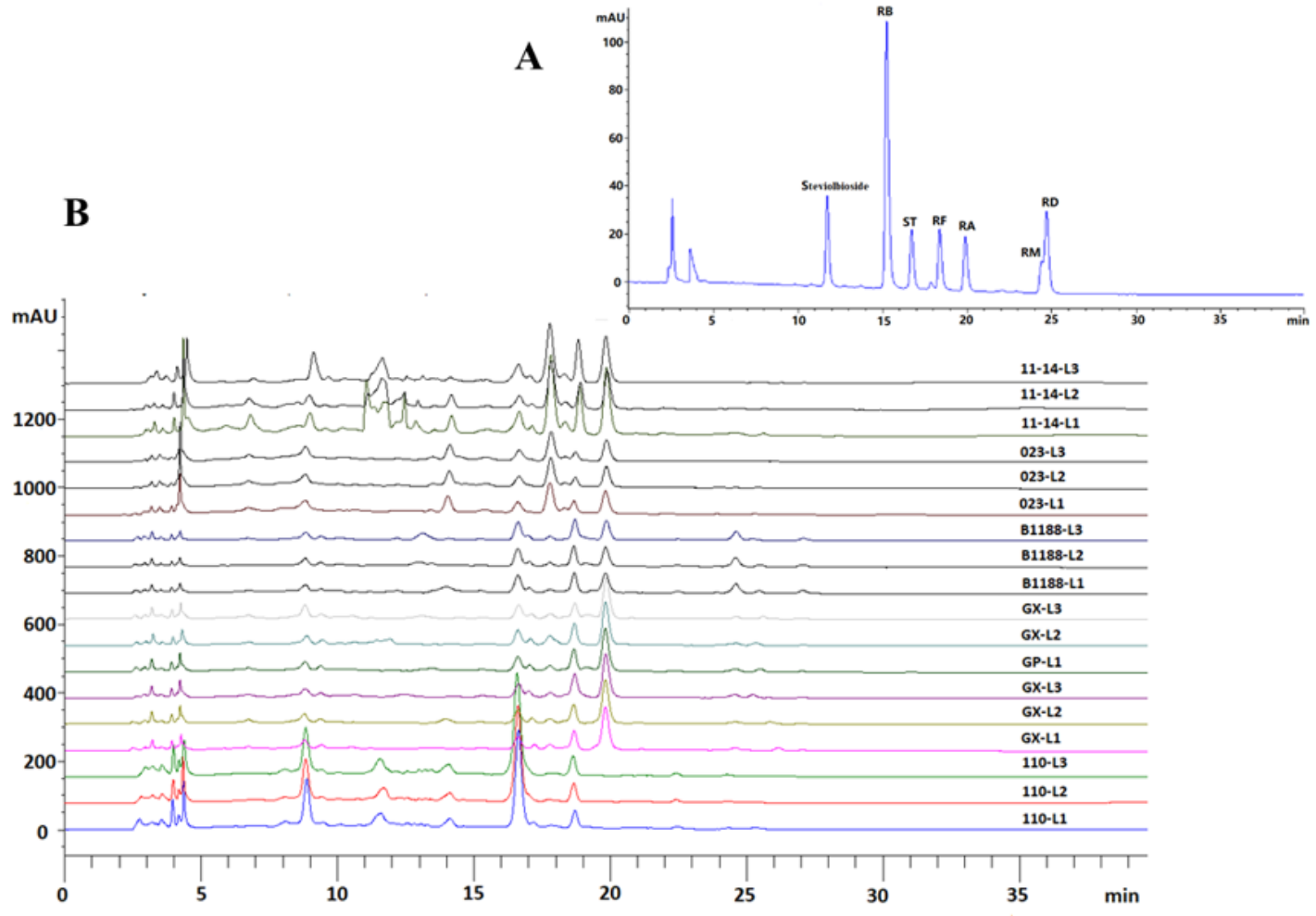

Figure 1

HPLC-UV profiles of the analytical samples and standards (steviolbioside, Reb B, ST, Reb F, Reb A, Reb D and Reb M). A. HPLC-UV profiles of the seven standards B. HPLC-UV profiles of the analytical samples

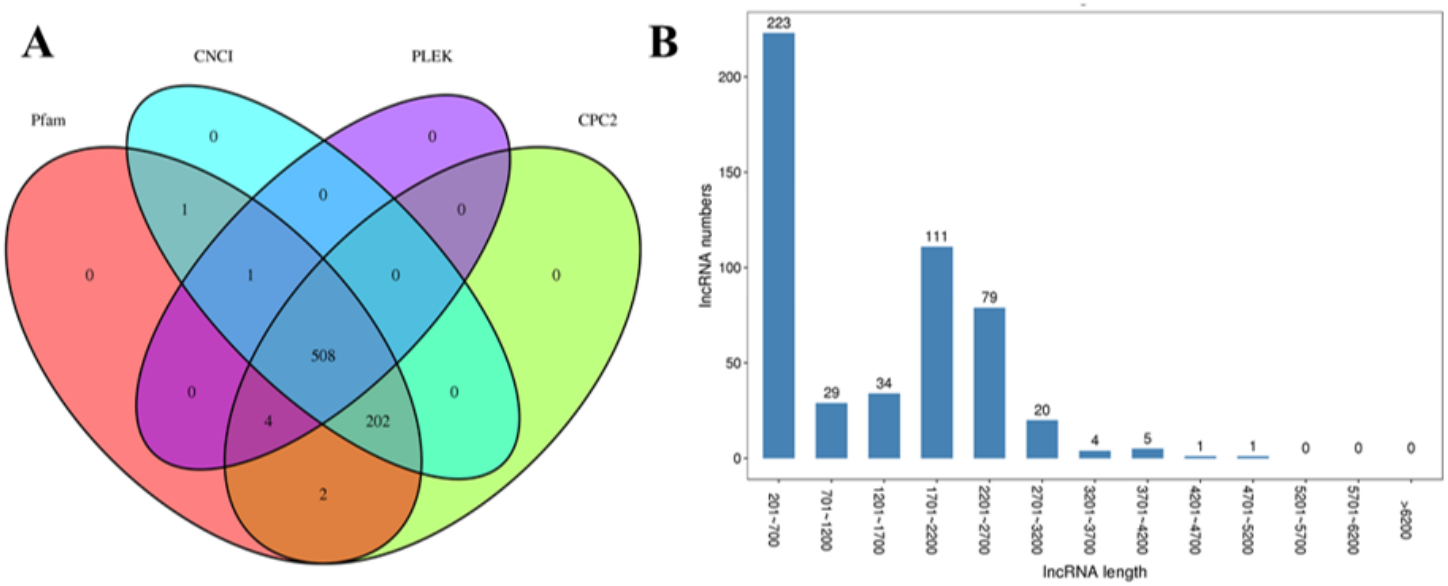

Figure 2

A. Noncoding Venn diagrams; B. LncRNA length distribution

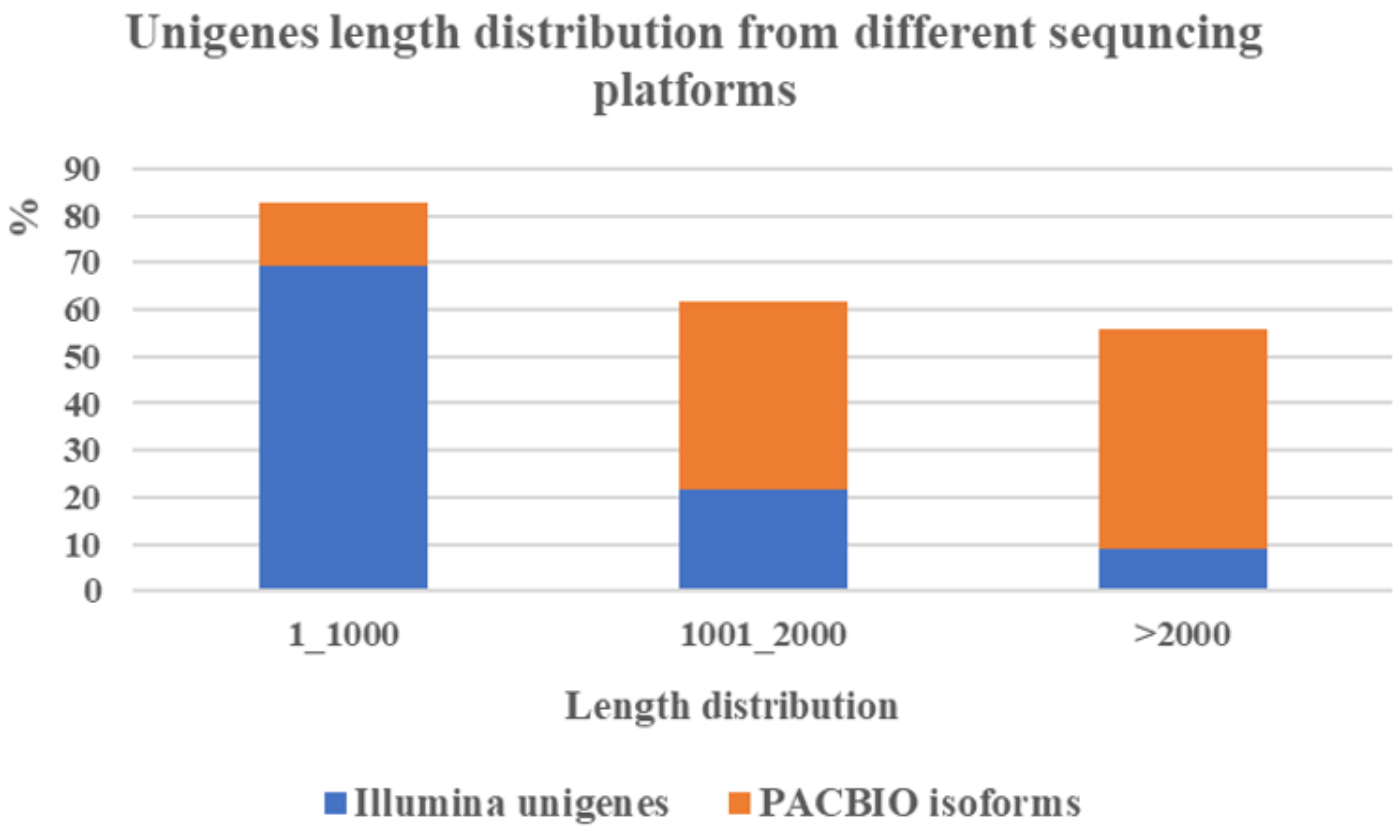

Figure 3

Comparison of unigene length distribution sequencing on different platforms.

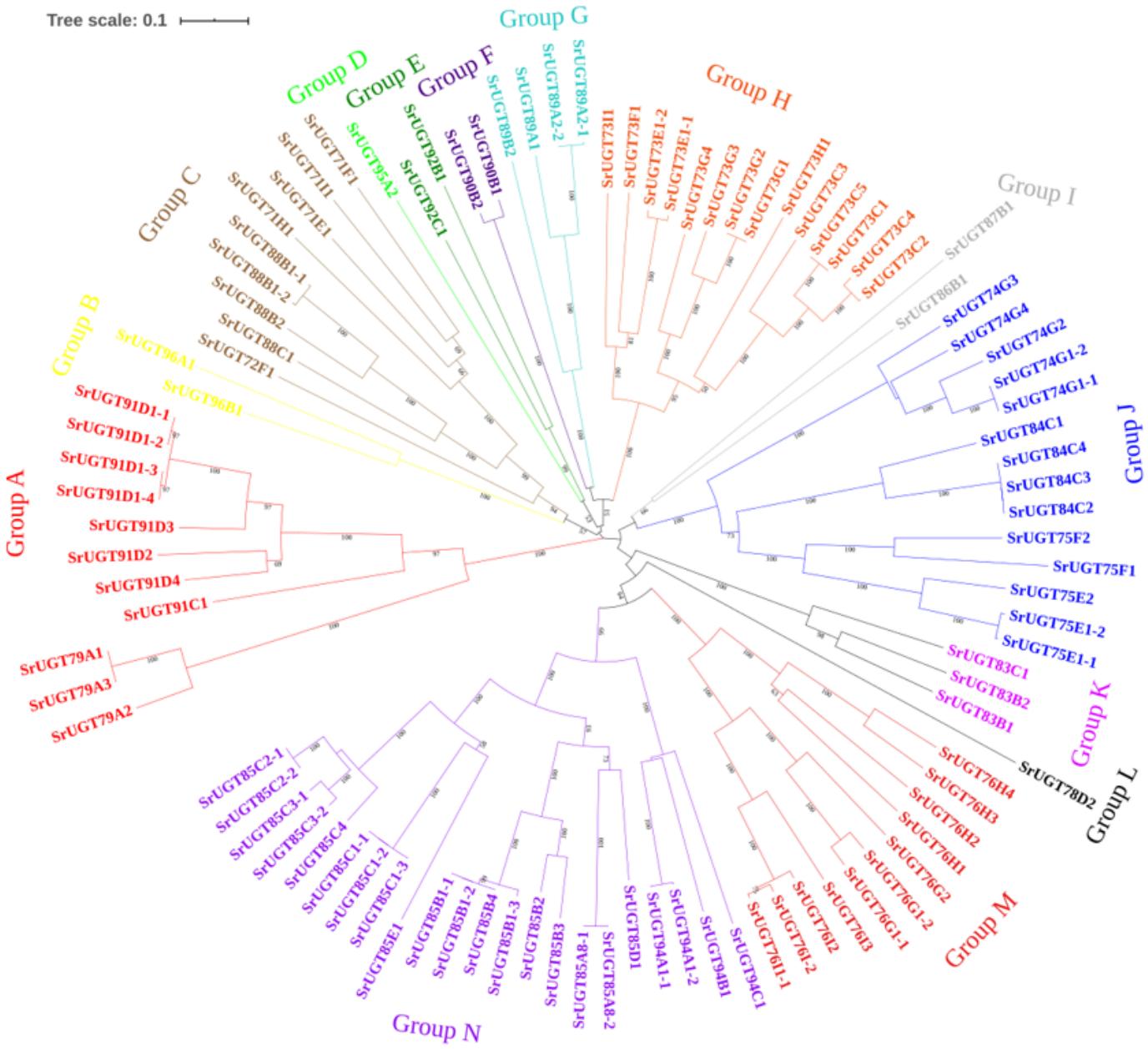

Figure 4

Phylogenetic analysis of the S. rebaudiana UGT superfamily shows 14 distinct groups, each with a bootstrap support greater than 90% in distance analysis excluding group I (66% bootstrap). The tree shown was derived by neighbour-joining distance analysis of the full-length amino acid sequence described in Figure S1. Distance bootstrap analyses consisted of 1000 replicates. Bootstrap values are listed as percentages of the replications, where values over 50% are indicated above the nodes.

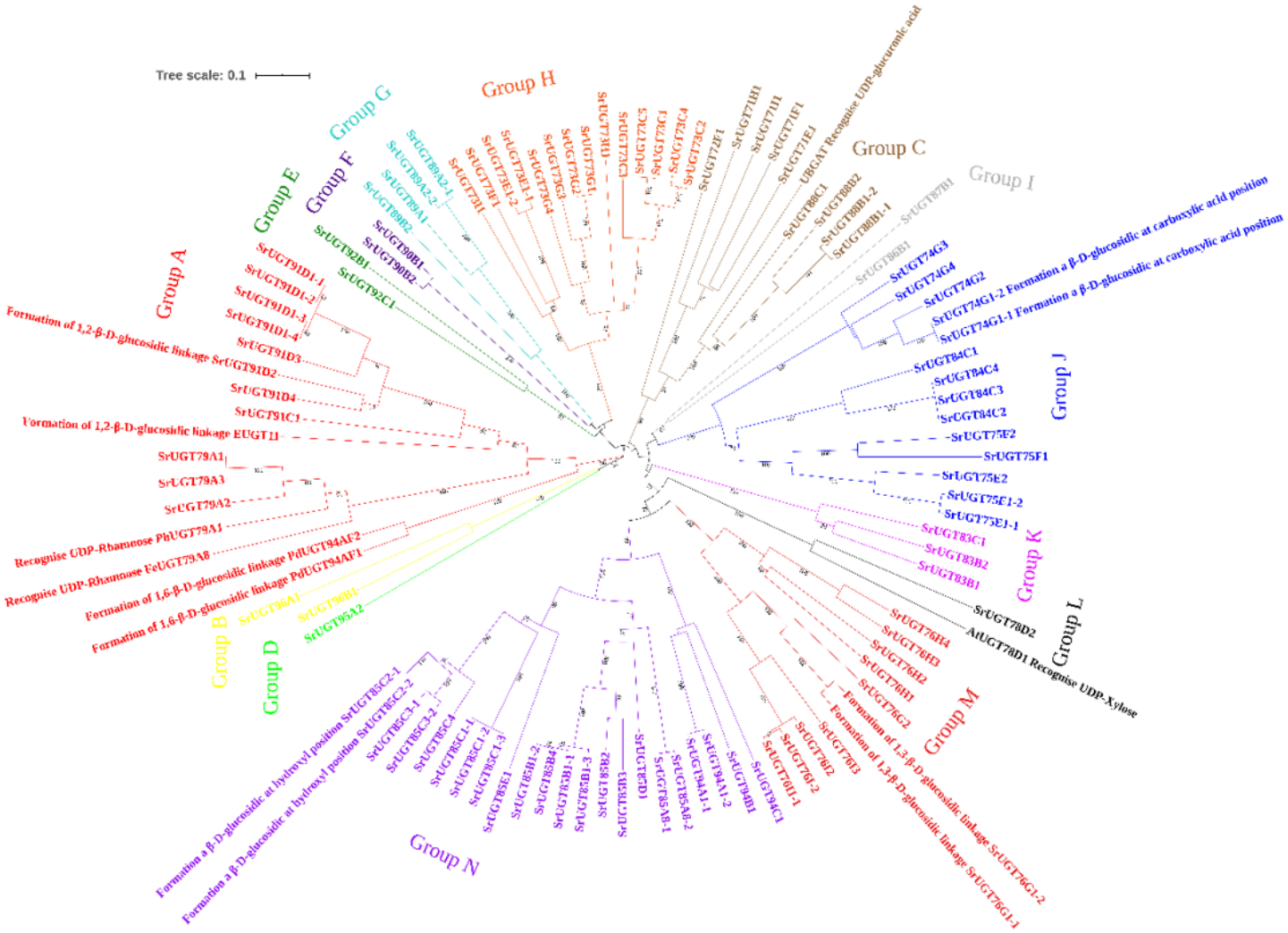

Figure 5

Composite phylogenetic tree. Seven UGTs from other plant species and four SrUGTs identified previously were also annotated in this tree. Distance bootstrap analyses consisted of 1000 replicates. Bootstrap values are listed as percentages of the replications, where values over 50% are indicated above the nodes.

Figure 6

Comparison of RNA-seq and qRT-PCR data. RNA-seq data and qRT-PCR quantification of changes in nine selected differentially expressed SrUGT genes in all sequenced materials. Transcript expression of 26S rRNA was used as an internal control. Blue represents the qRT-PCR data, and red represents the RNA-Seq data. The results showed that the expression levels of these analysed genes were similar between the qRT-PCR and transcriptome data.

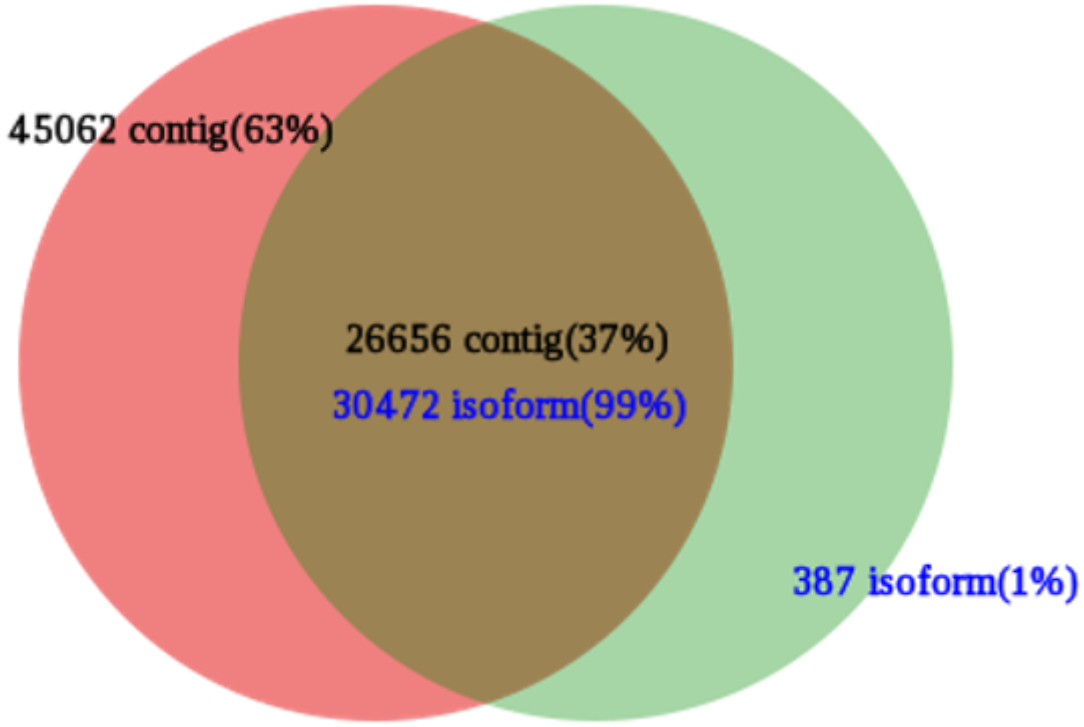

Figure 7

Intersection Venn of transcripts. Note: isoform is the number of transcripts sequenced from PACBIO, contig is the number of assembled transcripts sequenced from Illumina.

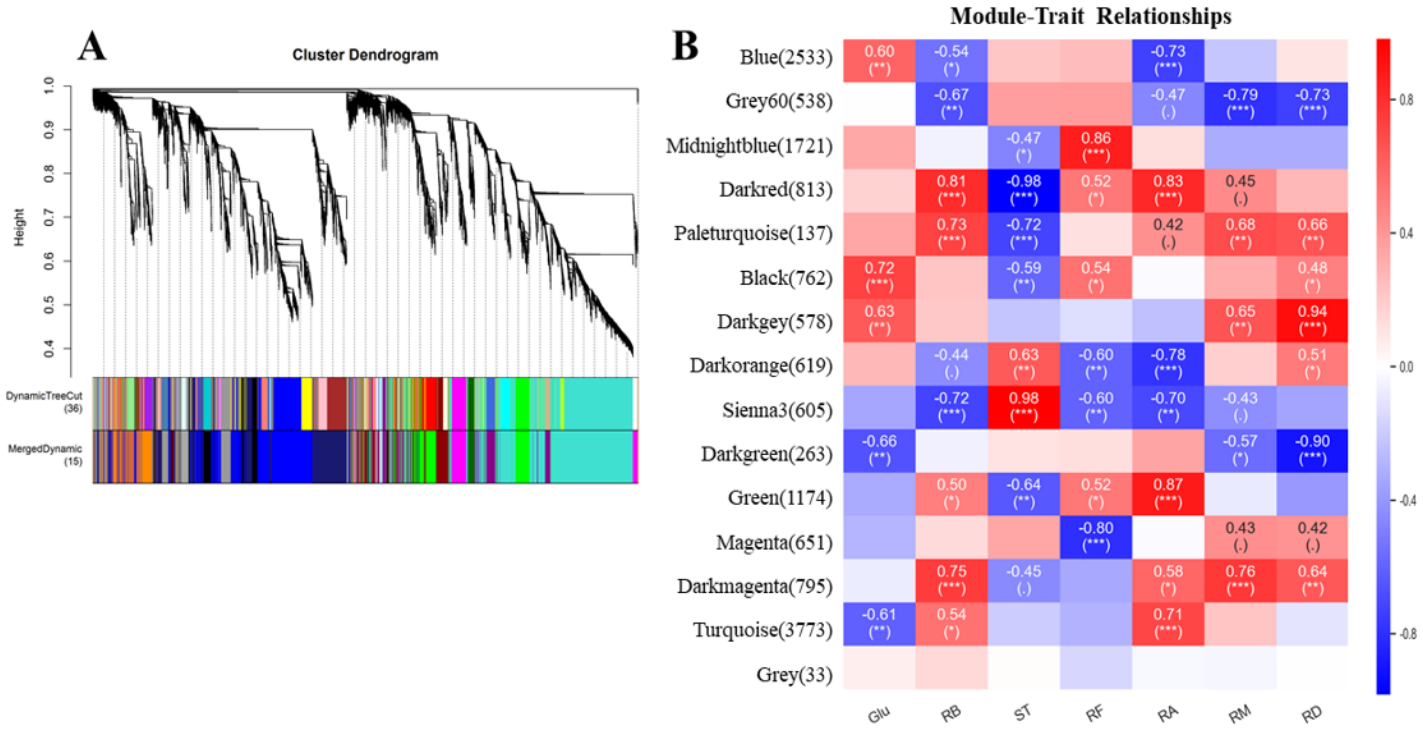

Figure 8

Network visualization plots. (A) Clustering dendrograms of genes with dissimilarity based on topological overlap together with assigned module colours. The same colour represents the same module, and the modules with certain correlations are merged into the same module. (B) Module-trait relationships. Each row corresponds to a module gene and column to a trait. Each cell contains the corresponding correlation and the significance. The table is colour-coded by correlation according to the colour legend.

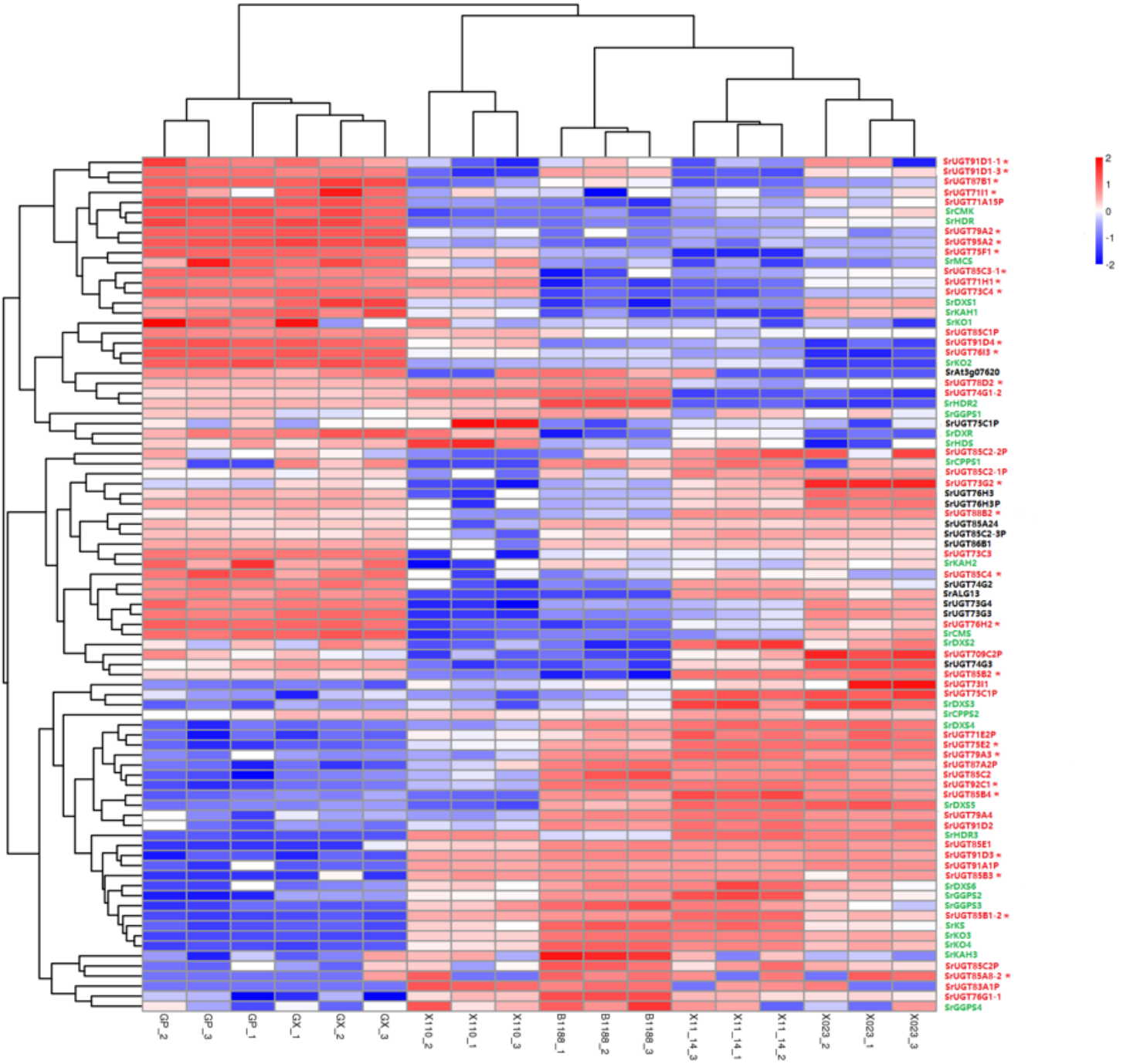

Figure 9

Heat map of the fifty-seven potential glucosyltransferase genes (in red and black) obtained from the weighted correlation network analysis and the genes (in green) known to be involved in steviol glycoside biosynthesis. Genes were clustered by expression patterns. Among the fifty-five glucosyltransferase

genes, forty-four SrUGTs (in red) are co-expressed with at least one upstream gene, including SrDXS, SrDXR, SrCMS, SrCMK, SrMCS, SrHDS, SrHDR, SrGGPS, SrCPPS, SrKS, SrKO and SrKAH. Of these genes, those having the P letter at the end were found to contain a single nucleotide substitution, insertion, or deletion, thereby causing interruption of ORFs. The presence of * means that further research will be performed in this study.

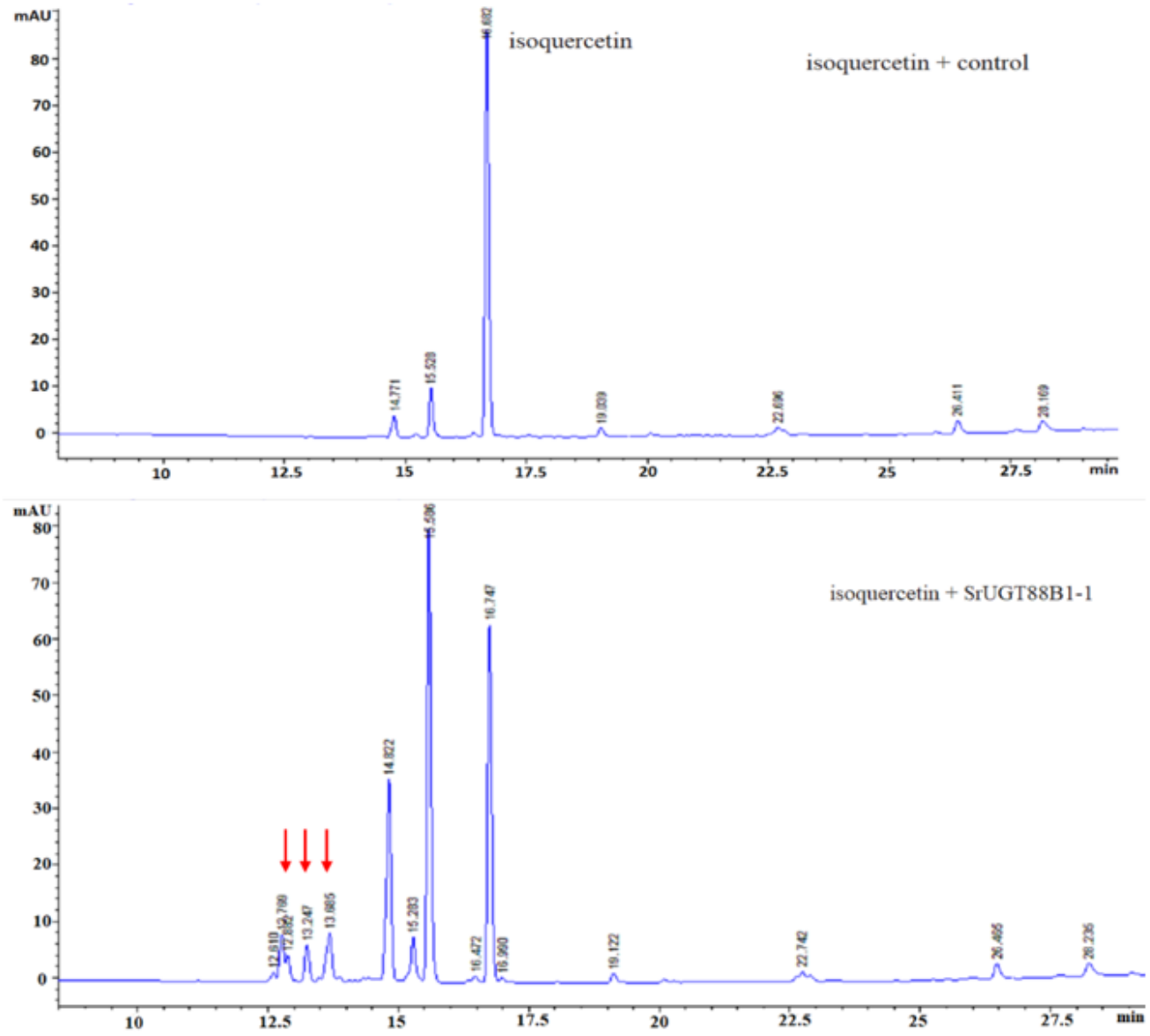

Figure 10

HPLC showing the peaks of the reaction products from SrUGT88B1-1 + isoquercetin and from the pET28a control vector + isoquercetin. The arrow indicates the three new compounds.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfiles.docx