Patil et al. BMC Genomics (2015) 16:520
DOI 10.1186/s12864-015-1730-y

BMC Genomics

RESEARCH ARTICLE Open Access

Soybean (*Glycine max*) SWEET gene family: insights through comparative genomics, transcriptome profiling and whole genome re-sequence analysis

Gunvant Patil[1], Babu Valliyodan[1], Rupesh Deshmukh[1], Silvas Prince[1], Bjorn Nicander[2], Mingzhe Zhao[1,4], Humira Sonah[1], Li Song[1], Li Lin[1], Juhi Chaudhary[1], Yang Liu[3], Trupti Joshi[3], Dong Xu[3] and Henry T. Nguyen[1*]

Abstract

Background: SWEET (*MtN3_saliva*) domain proteins, a recently identified group of efflux transporters, play an indispensable role in sugar efflux, phloem loading, plant-pathogen interaction and reproductive tissue development. The SWEET gene family is predominantly studied in *Arabidopsis* and members of the family are being investigated in rice. To date, no transcriptome or genomics analysis of soybean SWEET genes has been reported.

Results: In the present investigation, we explored the evolutionary aspect of the SWEET gene family in diverse plant species including primitive single cell algae to angiosperms with a major emphasis on *Glycine max*. Evolutionary features showed expansion and duplication of the SWEET gene family in land plants. Homology searches with BLAST tools and Hidden Markov Model-directed sequence alignments identified 52 SWEET genes that were mapped to 15 chromosomes in the soybean genome as tandem duplication events. Soybean SWEET (*GmSWEET*) genes showed a wide range of expression profiles in different tissues and developmental stages. Analysis of public transcriptome data and expression profiling using quantitative real time PCR (qRT-PCR) showed that a majority of the *GmSWEET* genes were confined to reproductive tissue development. Several natural genetic variants (non-synonymous SNPs, premature stop codons and haplotype) were identified in the *GmSWEET* genes using whole genome re-sequencing data analysis of 106 soybean genotypes. A significant association was observed between SNP-haplogroup and seed sucrose content in three gene clusters on chromosome 6.

Conclusion: Present investigation utilized comparative genomics, transcriptome profiling and whole genome re-sequencing approaches and provided a systematic description of soybean SWEET genes and identified putative candidates with probable roles in the reproductive tissue development. Gene expression profiling at different developmental stages and genomic variation data will aid as an important resource for the soybean research community and can be extremely valuable for understanding sink unloading and enhancing carbohydrate delivery to developing seeds for improving yield.

Keywords: SWEET, Effluxer, Sugar transport, Sink, Whole genome re-sequencing, Soybean

* Correspondence: nguyenhenry@missouri.edu
[1]National Center for Soybean Biotechnology and Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA
Full list of author information is available at the end of the article

Background

Photosynthesis fixes carbon in the leaves to make sugars as the primary transportable form of energy. Sugar production, status, and transport to the various tissues modulate the growth, productivity, and yield of plants [1]. In addition to their essential roles as substrates in carbon and energy metabolism, sugars also play an important role in signal transduction [1, 2]. In plants, sugars are accumulated in the form of simple sugars, carbohydrates, and starch. Stored sugars are then transported from leaves (source tissue) to the other plant parts (sink tissue) such as roots, modified leaves, and reproductive tissues (seeds). This transport from source to sink is modulated via phloem sap. Sucrose is synthesized in the cytosol and translocated to other non-photosynthetic tissues for direct metabolic use or for conversion to starch. Allocation of sucrose is facilitated by both short-distance transport systems and long-distance transport systems [3]. Short distance transport is achieved at the intra-cellular and inter-cellular levels, where sucrose is transported via diffusion/protoplasmic streaming and plasmodesmata, respectively [4, 5]. It then moves from cell to cell via plasmodesmata until it reaches the phloem parenchyma cells, and in the phloem parenchyma cells, processes related to long-distance transport initiate [6, 7]. Among the sugars, only a few are allocated to the phloem long-distance transport system and sucrose is the main form of carbon found in the phloem tissue followed by polyols, raffinose, etc. [8, 9]. Of the many different sugars found in plants, it is mainly sucrose that is transported in the phloem, where it is the most abundant carbonaceous compound [8].

The amount of sucrose available for transport to the sink tissues is very crucial for plant development [8, 10]. Metabolite transport efficiency influences photosynthetic productivity by relieving product inhibition and contributes to plant vigor by controlling source/sink relationships and biomass partitioning. The sucrose transport is controlled or facilitated by SUT (sucrose transporter) [11–13] and SWEET (sucrose effluxer) proteins [14–16]. SUT has been widely studied in many plant species [4, 11–14, 17, 18]. SUT proteins are expressed at low levels and display saturable sucrose transport kinetics, suggesting that additional transport proteins are responsible for sucrose allocation across the membrane [6]. The milestone efforts that identified the sucrose effluxer was led by Chen et. al. (2010) [15]. They identified the role of the SWEET (**S**ugars **W**ill **E**ventually be **E**xported **T**ransporters) gene family as sucrose effluxers based on their role in transporting glucose molecules across a membrane. SWEET proteins contains a *MtN3_slv* transmembrane domain that is essential for the maintenance of animal blood glucose levels, plant nectar production, and plant seed and pollen development [19, 20]. The first member of the SWEET family, *MtN3*, was identified as a nodulin-specific EST in the leguminous plant *Medicago truncatula* [21], and *MtN3_slv* was identified as an embryonic salivary gland specific gene in drosophila [22]. SWEET proteins function as uniporters, facilitate diffusion of sugars across cell membranes, and mediate sucrose efflux from putative phloem parenchyma into the phloem apoplasm [23–25]. In Arabidopsis, members of the SWEET gene family, *AtSWEET11* and *–12* were localized to the plasma membrane of the phloem parenchyma and are the main facilitators of sucrose flux. Mutations in *AtSWEET11, –12* genes led to defective phloem loading without affecting the phenotype [26]. Using optical sucrose sensors, SWEET proteins were identified as assisting movement of sucrose across cell membranes in preparation for long-distance transport. SWEET proteins are expressed in phloem parenchyma cells and are key to the export of sucrose from leaves [26].

SWEET transporters have diverse physiological roles and are essential for the maintenance of animal blood glucose levels, plant nectar production, and plant seed and pollen development [15, 23]. Arabidopsis *AtSWEET8* is essential for pollen viability, and the rice homologous *OsSWEET11* and *OsSWEET14* are specifically exploited by bacterial pathogens for virulence by means of direct binding of a bacterial effector to the SWEET promoter [27, 28]. Bacterial and fungal symbionts/pathogens induce the expression of different SWEET genes by secreting the effector protein that binds and activates SWEET genes, indicating that the sugar efflux function of SWEET transporters is targeted and hijacked by pathogens and symbionts for nutritional gain [5, 6, 15, 28, 29].

The sink organs, especially developing seeds which are mainly heterotrophic, depend on nutrients from their parent plants [30, 31]. Early development of the embryo is controlled by the maternal tissue and then during maturation it is controlled by the filial tissues [32]. Phloem unloading in most of the sink tissues follows symplasmic routes [30, 33]. In many dicot seeds, e.g. legumes [32] and Arabidopsis [31], the filial tissues are symplasmically isolated/interrupted by apoplast from the phloem in the maternal seed tissue. Transport of sucrose from phloem to the filial tissue is associated with the expression of sugar transporters, localized to the plasma membranes of filial cells. [5, 25, 33–36]. Ludewig et. al. [37] and Braun [24] have reviewed and discussed role of the SWEET family transporter as putative facilitator of phloem unloading or as the transporter mediating diffusion of sucrose in sink tissue. Similarly, it has been proposed that enhancing nutrient flow to the developing endosperm and embryo by overexpressing SWEET genes along with cell wall invertase and hexose symporter genes at seed maternal-filial interface can increase the seed yield [5, 38]. In *M. domestica*, the SWEET genes,

including other sugar transporter genes, are involved in sugar accumulation in sink tissue and the concentration of sugars were positively correlated with the SWEET gene expression [39].

To date SWEET genes are well studied in Arabidposis [15, 26] and rice [20, 40] but no genome-wide exploration and characterization of the SWEET gene family has been performed in soybean. Studying the sucrose efflux system across different species and genera will lead us to understand the evolutionary aspects of the SWEET gene family. In this study, we first collected the SWEET gene family in a number of plant species, then focused on soybean where 52 putative SWEET genes were identified. The publicly available transcriptome datasets were explored and the expression pattern of 23 genes were analyzed using qRT-PCR in reproductive tissues. The wealth of whole genome re-sequencing resources in soybean provided an opportunity to explore natural variation in the soybean SWEET genes. The data presented here lays the foundation for further investigations into the biological and physiological processes of SWEET genes in soybean.

Results

Identification of SWEET genes in soybean and other species

To find soybean SWEET homologues, BLAST and PFAM [41] searches were performed using *Arabidposis* and rice SWEET genes. This led to the identification of 52 genes with high homology (Fig. 1, Additional file 1). This is far higher than in the other 24 species used in this study. A number of genes with lower homology to SWEET were also found, but were not studied further. The 52 soybean SWEET genes identified in our study were designated as *GmSWEET1* to *GmSWEET*52. Similarly SWEET genes in other species were extracted from the Plaza comparative genomics platform [42] using BLASTN and BLASTP searches, 444 SWEET genes (including 33 outliers) were predicted across 25 genomes (Additional file 1). The details about other parameters, including nucleic acid and protein sequences, are provided in Table 1 and Additional file 1.

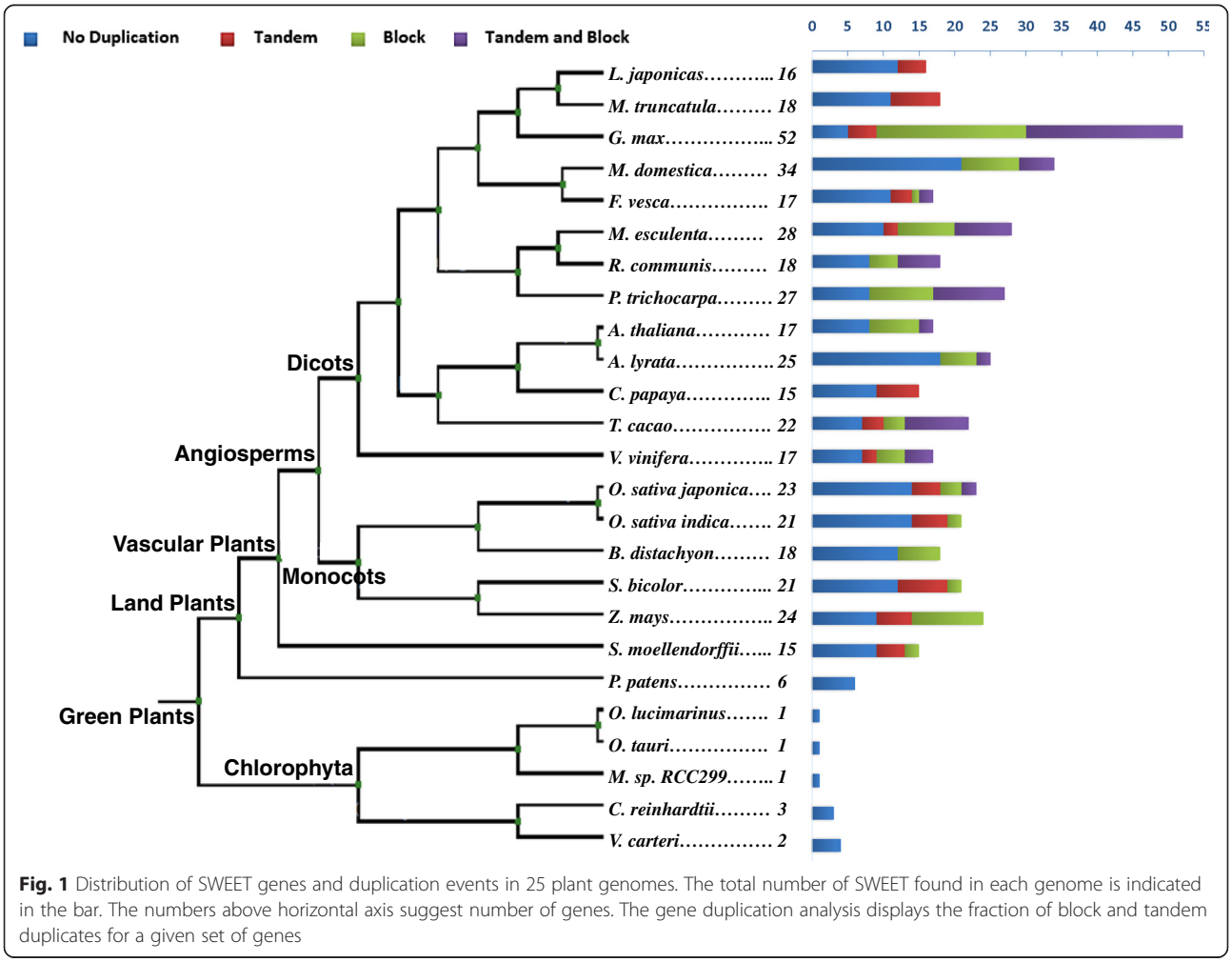

Fig. 1 Distribution of SWEET genes and duplication events in 25 plant genomes. The total number of SWEET found in each genome is indicated in the bar. The numbers above horizontal axis suggest number of genes. The gene duplication analysis displays the fraction of block and tandem duplicates for a given set of genes

Table 1 List of 52 soybean SWEET genes and their sequence details (aa- amino acid)

Name	Glyma ID V1.0	Protein aa	Chr. No.	Intron	TMs	mRNA coordinates		Arabidopsis Orthologs
						Start	End	
GmSWEET1	*Glyma02g09710*	262	2	5	7	7672242	7674181	AtSWEET9
GmSWEET2	*Glyma03g36790*	316	3	8	7	43645675	43647218	AtSWEET9
GmSWEET3	*Glyma03g39430*	155	3	4	4	45507339	45508526	AtSWEET16/17
GmSWEET4	*Glyma04g37510*	259	4	5	7	43916099	43918975	AtSWEET10
GmSWEET5	*Glyma04g37520*	283	4	5	7	43926535	43929580	AtSWEET10
GmSWEET6	*Glyma04g37530*	277	4	4	6	43938391	43940184	AtSWEET11/12/13/14
GmSWEET7	*Glyma04g41680*	175	4	4	5	47528111	47529572	AtSWEET3
GmSWEET8	*Glyma04g42040*	248	4	5	7	47812561	47815829	AtSWEET1
GmSWEET9	*Glyma05g02070*	226	5	4	6	1492949	1494364	AtSWEET4
GmSWEET10	*Glyma05g25180*	283	5	3	7	31313674	31315182	AtSWEET15
GmSWEET11	*Glyma05g38340*	258	5	5	7	41712591	41715252	AtSWEET10
GmSWEET12	*Glyma05g38350*	276	5	6	6	41723915	41726765	AtSWEET11/12/13/14
GmSWEET13	*Glyma06g12740*	259	6	5	7	9947277	9952565	AtSWEET1
GmSWEET14	*Glyma06g13110*	255	6	5	7	10255499	10257429	AtSWEET3
GmSWEET15	*Glyma06g17520*	310	6	5	7	13868324	13870606	AtSWEET11/12/13/14
GmSWEET16	*Glyma06g17530*	261	6	5	7	13887645	13890535	AtSWEET10
GmSWEET17	*Glyma06g17540*	259	6	5	7	13901720	13904519	AtSWEET10
GmSWEET18	*Glyma06g21570*	244	6	7	5	18131992	18134116	AtSWEET16/17
GmSWEET19	*Glyma06g21640*	192	6	3	4	18206442	18207648	AtSWEET16/17
GmSWEET20	*Glyma08g01300*	295	8	5	7	771448	773996	AtSWEET11/12/13/14
GmSWEET21	*Glyma08g01310*	255	8	5	7	781403	783956	AtSWEET10
GmSWEET22	*Glyma08g02890*	274	8	4	7	1977779	1981412	AtSWEET15
GmSWEET23	*Glyma08g08200*	260	8	5	7	5861243	5863837	AtSWEET15
GmSWEET24	*Glyma08g19580*	281	8	5	7	14793461	14795629	AtSWEET15
GmSWEET25	*Glyma08g47550*	272	8	5	7	46378609	46380649	AtSWEET15
GmSWEET26	*Glyma08g47560*	274	8	5	7	46385711	46388217	AtSWEET15
GmSWEET27	*Glyma08g48280*	224	8	2	6	46926095	46926877	AtSWEET9
GmSWEET28	*Glyma09g04840*	245	9	5	7	3652169	3657426	AtSWEET16/17
GmSWEET29	*Glyma12g36300*	236	12	5	7	39418338	39419741	AtSWEET2
GmSWEET30	*Glyma13g08190*	256	13	5	7	8569038	8571904	AtSWEET3
GmSWEET31	*Glyma13g09140*	249	13	5	7	10118236	10121745	AtSWEET1
GmSWEET32	*Glyma13g10560*	258	13	4	7	12437340	12439924	AtSWEET6/7
GmSWEET33	*Glyma13g23860*	246	13	5	6	27171336	27175275	AtSWEET4
GmSWEET34	*Glyma13g33950*	236	13	5	7	35599596	35602840	AtSWEET2
GmSWEET35	*Glyma14g17810*	181	14	7	5	19873917	19875274	AtSWEET9
GmSWEET36	*Glyma14g27610*	250	14	5	7	33859148	33862494	AtSWEET1
GmSWEET37	*Glyma14g30740*	247	14	6	7	37447737	37449539	AtSWEET3
GmSWEET38	*Glyma14g30940*	255	14	5	7	37715242	37718176	AtSWEET3
GmSWEET39	*Glyma15g05470*	250	15	5	7	3856854	3858704	AtSWEET15
GmSWEET40	*Glyma15g16030*	246	15	5	7	12350866	12354851	AtSWEET16/17
GmSWEET41	*Glyma15g27530*	262	15	5	6	29768360	29770247	AtSWEET2
GmSWEET42	*Glyma15g27750*	236	15	5	7	30321666	30324324	AtSWEET2
GmSWEET43	*Glyma17g09840*	227	17	5	6	7343156	7345793	AtSWEET4

Table 1 List of 52 soybean SWEET genes and their sequence details (aa- amino acid) *(Continued)*

GmSWEET44	*Glyma18g53250*	263	18	5	7	61559244	61561078	AtSWEET9
GmSWEET45	*Glyma18g53930*	269	18	5	7	62184272	62186255	AtSWEET15
GmSWEET46	*Glyma18g53940*	272	18	5	7	62193884	62196748	AtSWEET15
GmSWEET47	*Glyma19g01270*	232	19	4	7	880980	884031	AtSWEET4
GmSWEET48	*Glyma19g01280*	247	19	5	6	886189	893470	AtSWEET4
GmSWEET49	*Glyma19g42040*	308	19	5	7	48123304	48127034	AtSWEET16/17
GmSWEET50	*Glyma20g01890*	160	20	3	1	1422950	1425797	AtSWEET16/17
GmSWEET51	*Glyma20g16160*	257	20	4	6	22458358	22460991	AtSWEET6/7
GmSWEET52	*Glyma20g21060*	213	20	4	4	29990001	29991796	AtSWEET16/17

Soybean SWEET genes are highly conserved and points to duplication events in higher plants

Comparative genomics of SWEET genes were performed using 25 plant genomes encompassing monocots, dicots and lower plants with subsequent focus on the soybean SWEET family. According to a database of conserved protein families (PFAM), *MtN3-like* clan (http://pfam.xfam.org/clan/MtN3-like) contains five subfamilies: *MtN3_slv* (PF03083), PQ-loop (PF04193), MPC (PF03650), ER Lumen Receptor (PF00810), and Lab-N (PF07578). The SWEET genes belongs to *MtN3_slv* subfamily and serve function in sugar transport whereas other proteins have different roles, for example PQ-loop subfamily involved in amino acid transport [43]. SWEET gene (*MtN3_slv*) homologues from algae, moss and higher plants were collected from Genbank and Plaza 2.5 and 3.0 comparative genomics platforms [42]. Genome-wide distribution of the SWEET gene family showed that the unicellular plants and blue green algae have fewer copies (1–4) of SWEET genes, followed by 6 and 15 genes in *Physcomitrella patens* (non-vascular) and *Selaginella moellendorffii* (vascular) from lower plant group, respectively (Fig. 1).

To better understand the evolutionary relationship between different plant SWEET (*MtN3_slv*) homologues, we constructed a phylogenetic tree using 173 SWEET genes from 13 species representing major plant groups (Fig. 2, Amino acid sequences see Additional file 2). These 13 plant species represent; dicots (*Glycine max, Medicago truncatula, Vitis vinifera, Arabidopsis thaliana*), monocots (*Oryza sativa*, *Zea mays*), bryophytes (*Physcomitrella patens*, *Selaginella moellendorffii*), and algae (*Ostreococcus lucimarinus Ostreococcus tauri, Micromonas sp. RCC299, Chlamydomonas reinhardtii, Volvox carteri*). The phylogenetic clustering between different plant species reveal the evolutionary relationship among plant SWEET proteins. Four major clades were perceived, in which both monocots and dicots were distributed between clades I-III. The algal species were observed in clade number IV and the bryophytes (*P.patens* and S. *moellendorffii*) were predominantly observed in clade I. Interestingly, four algal species, those of the unicellular chlorophyta group (*O. lucimarinus*, *O. tauri, M. sp.RCC299, C. reinhardtii*) contain only 7- transmembrane domains (TMs) and not 3-TMs, which led us to speculate that the multicellular plants (bryophytes and flowering plants) might have acquired 3-TMs from symbiotic bacteria through horizontal gene transfer or might have evolved through internal duplication of 3-TMs within the gene.

The phylogeny of the soybean SWEET genes was compared to the Arabidopsis and rice SWEET genes since they have been functionally characterized and their duplication events represented by a whole-genome duplication. The lineage-specific arrangement of SWEET genes proposes that the genes may be expanded and then diversified after the monocot and dicot division. Soybean contains the highest number (52) of SWEET homologues as compared to other plant species included in the present study. To gain further insight into the structural diversity of *GmSWEET* genes, we compared intron/exon organization in the coding sequences of paralog pairs and found that most of the paralogs shared similar gene organization, consistent with the phylogenetic analysis (Additional file 3).

Soybean SWEET gene family expansion

The phylogeny of the SWEET genes points to several duplication events. Out of 411 SWEET genes across 25 genomes, 56 tandem genes, 95 block duplication events, 72 genes were found to be both tandem and block duplication events (Fig. 1). The multiple sets of SWEET genes were first appeared in *S. moellendorffii* through duplication events. The non-vascular plant group (Chlorophyta and *P. patens*) did not show any gene duplication events. In soybean, 52 SWEET genes were mapped to 15 chromosomes and a majority were distributed in the more gene-dense euchromatic region near the chromosome ends (Fig. 3). The genes and clusters showed random distribution among the chromosomes. Chromosome numbers 2, 9, 12, and 17 contain only one SWEET gene, while chromosome 8 contains eight, the maximum number of SWEET genes per chromosome. It is known that

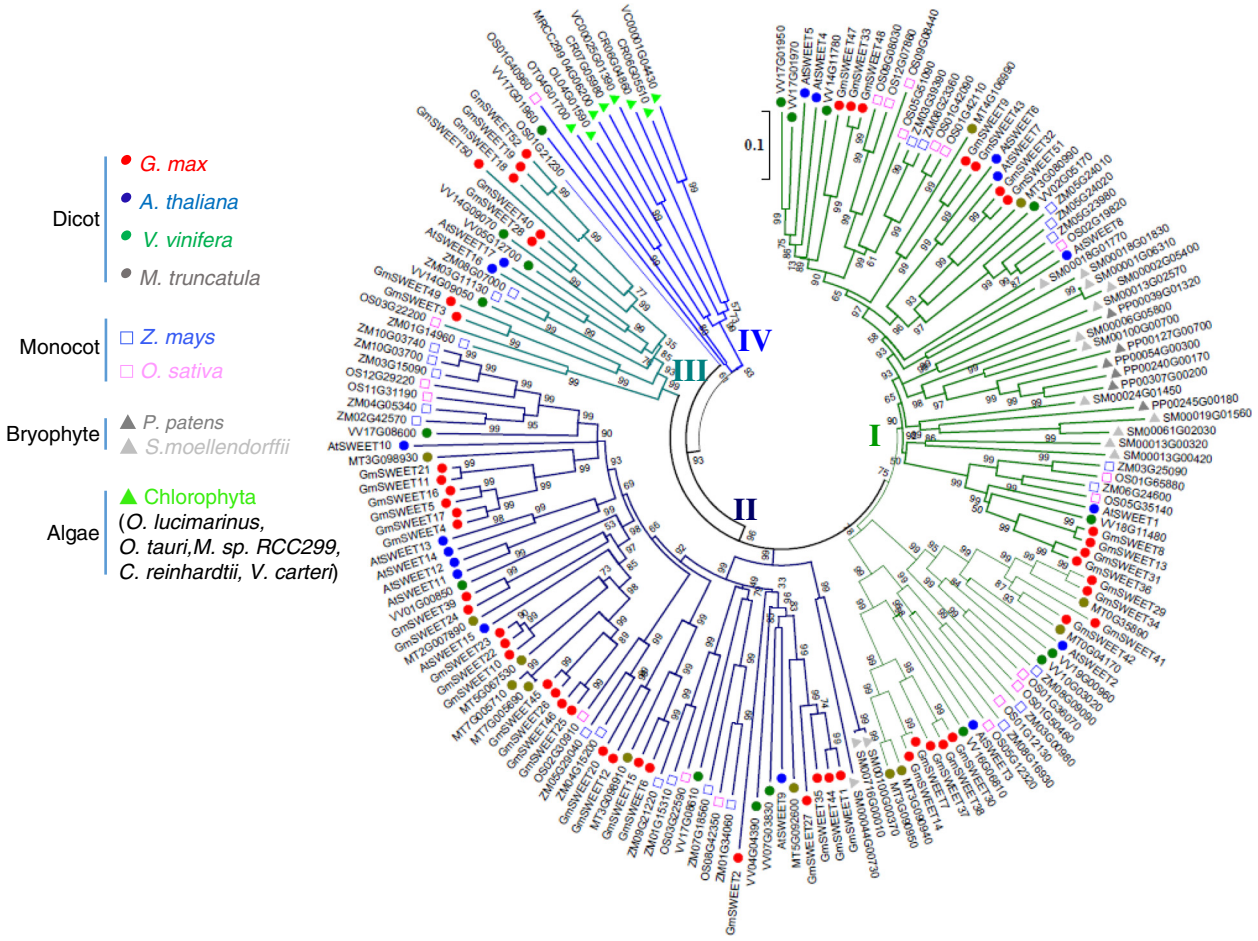

Fig. 2 Phylogenetic relationship of SWEET gene family proteins in 13 different species. The phylogenetic tree was built using the neighbor-joining (NJ) method implemented in MEGA5.1. The roman numerals (I – IV) indicated with different colors, represents the clades associated with higher and lower plant groups. The numbers at the nodes represent bootstrap percentage values based on 1000 replications. Genes from each species are marked with different bullet point colors

polyploidy is a crucial force in plant evolution, and many angiosperms have experienced one or more episodes of polyploidization which subsequently resulted in gene duplication within the gene family [44, 45]. Soybean paralogs within a gene family were derived from genome duplications that occurred approximately 130 million years ago (MYA) (before the origin of rosids), 59 MYA (during legume genome duplication), and 13 MYA (duplication in the Glycine lineage) and nearly 75 % of the genes are present in multiple copies [44, 46]. In soybean, 21 *GmSWEET* sister pairs were identified with higher bootstrap values (<90 %) and the duplication of genes in soybean resulted in gene family expansion. Interestingly, we found clusters of five genes (*GmSWEET* 4 to 8 and *GmSWEET* 13 to 17) that were tandemly duplicated between chromosome 4 and 6. Similar tandem duplication clusters were observed between chromosome 5 and 8 and chromosome 8 and 18 (Fig. 3). The synonymous substitution rates (Ks), the non-synonymous substitution rates (Ka) and the Ka/Ks ratio for the 21 duplicated gene pairs revealed high similarities in their coding sequence alignments. The Ks values of these 21 genes ranged from 0.03 for gene pair *Glyma05G02070/Glyma17G09840* to 0.18 for pair *Glyma04G37520/Glyma06G17530* with an average Ks of 0.105 (Table 2), which is consistent with genes that emerged from the most recent genome duplication event 13 MYA [46, 47]. The history of selection performed on coding sequences can be measured by the Ka/Ks ratio and can be used to identify pairwise combinations of genes, where encoded proteins may have changed function [48]. Ka/Ks < 1 indicates that those genes underwent a purifying (stabilizing) selection and Ka/Ks > 1 at specific sites indicate genes that are under positive selection or Darwinian selection [47]. Table 2 summarizes the Ka/Ks for 21 duplicated pairs, in which 20 pairs were less than 0.9, indicating purifying selection and one pair (*Glyma05G02070/Glyma17G09840*) had a value of 1.79 indicating the positive selection. Based on

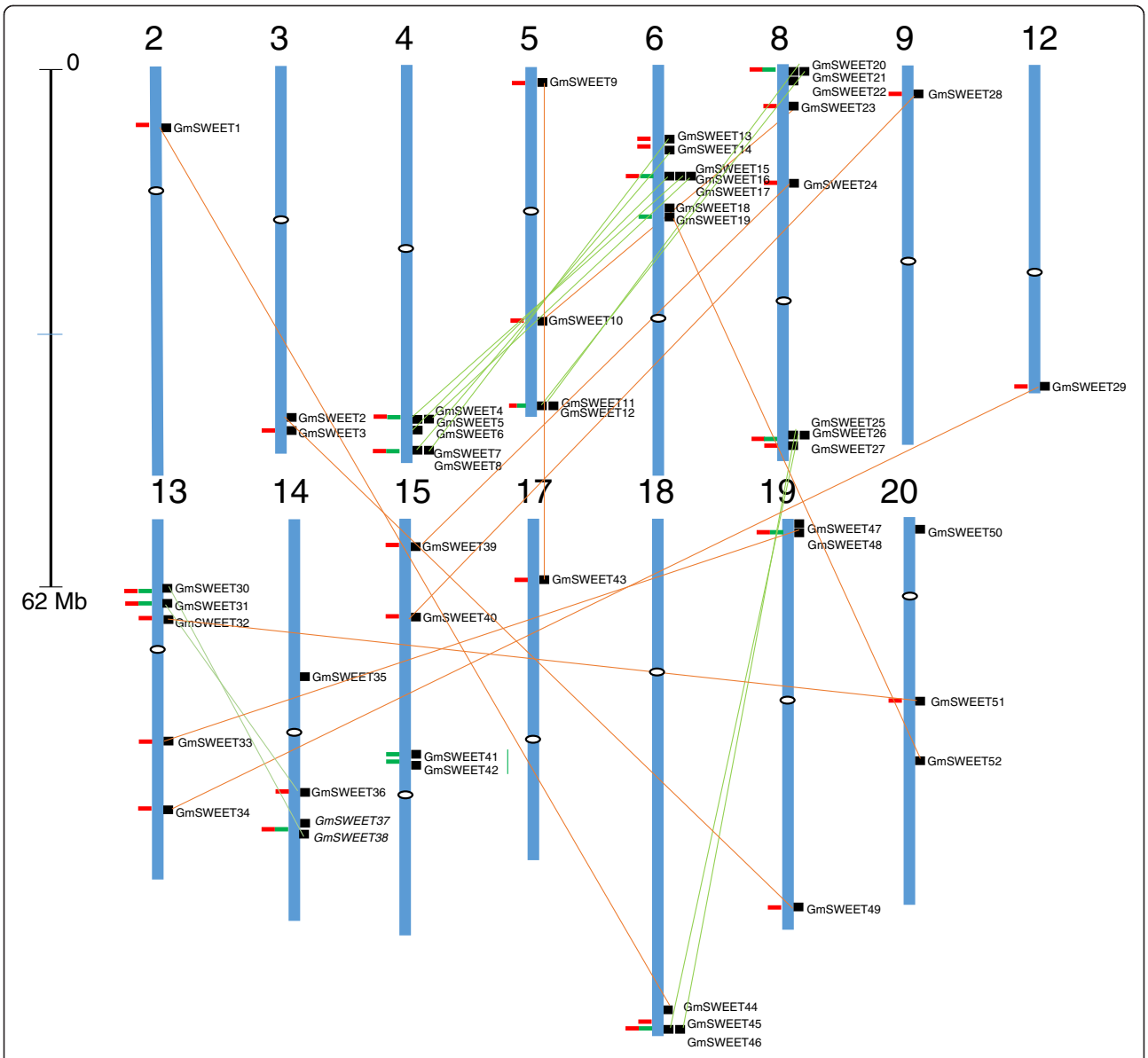

Fig. 3 Chromosomal locations of soybean SWEET genes. The 52 SWEET genes were mapped to the 15 out of 20 chromosomes. Black boxes represent the gene position on the chromosome. The data used to generate the schematic diagram of the genome-wide chromosome organization was obtained from Phytozome and SoyKB genome browsers. Tandem and block duplications are marked with bold green and red boxes, respectively. Homologues were connected by orange (non-clustered genes) and light green (clustered genes) lines. Black scale line represents the length of chromosome. White dots on each chromosome represents centromere position

the divergence rate of $\lambda = 6.1\times10^{-9}$ proposed for soybean [49], 20/21 SWEET paralogous pairs were estimated to have occurred between 4.95 to 14.9 MYA, except one pair at 2.88 MYA.

Conserved domains

The typical SWEET protein contains seven TM helices consisting of two tandem repeats of 3-TM units separated by a single TM unit [43]. Prokaryotes have homologues with only 3-TM units (semiSWEETs), which assemble into multiple 3-TM unit complexes to mediate sucrose transport [43, 50]. On the other hand, eukaryotes have both 7-TM and 3-TM SWEET genes. The eukaryotic 7-TMs have evolved by internal duplication of the 3-TMs [43] (see Fig. 4 for overall structural relationship of the sub-types). To understand the conservation of different domain within the gene family the protein sequences were aligned. On average, SWEET proteins in plants contain 5 exons that form a protein with an average of 248 amino acids. We found that out of 411 SWEETs, 140 were semi-SWEET genes, each either missing the first or the second 3-TM domain, or they were present only in a partial form

Table 2 Identification of substitution rates for homologues *GmSWEET* genes

Gene ID	No. of syn sites	No. of non-syn sites	Syno. substitution rate (Ks)	Non-syn substitution rate (Ka)	Ka/Ks	Duplication Date (MYA)
GmSWEET29	*214.9*	*490.1*	*0.0982*	*0.0229*	*0.2329*	8.05
GmSWEET34						
GmSWEET41	*179.6*	*489.4*	*0.1137*	*0.098*	*0.8627*	9.32
GmSWEET42						
GmSWEET7	*156.9*	*365.1*	*0.0983*	*0.011*	*0.1121*	8.06
GmSWEET14						
GmSWEET30	*210.8*	*551.2*	*0.1069*	*0.0225*	*0.2103*	8.76
GmSWEET38						
GmSWEET8	*207.2*	*533.8*	*0.1056*	*0.0112*	*0.1063*	8.66
GmSWEET13						
GmSWEET31	*209.6*	*534.4*	*0.0977*	*0.0346*	*0.3548*	8.01
GmSWEET36						
GmSWEET1	*218.2*	*567.8*	*0.1049*	*0.0259*	*0.2464*	8.60
GmSWEET44						
GmSWEET6	*222.4*	*614.6*	*0.09*	*0.0442*	*0.4912*	7.38
GmSWEET15						
GmSWEET12	*215.2*	*492.8*	*0.1059*	*0.0318*	*0.3001*	8.68
GmSWEET20						
GmSWEET11	*191.7*	*570.3*	*0.0866*	*0.0122*	*0.1408*	7.10
GmSWEET21						
GmSWEET4	*224.1*	*549.9*	*0.1265*	*0.0318*	*0.2513*	10.37
GmSWEET17						
GmSWEET5	*220.5*	*556.5*	*0.1823*	*0.0396*	*0.2172*	14.94
GmSWEET16						
GmSWEET24	*182.1*	*537.9*	*0.0955*	*0.0415*	*0.4342*	7.83
GmSWEET39						
GmSWEET10	*239.6*	*579.4*	*0.1145*	*0.0287*	*0.251*	9.39
GmSWEET23						
GmSWEET26	*234.6*	*569.4*	*0.0605*	*0.0108*	*0.178*	4.96
GmSWEET45						
GmSWEET25	*213.8*	*599.2*	*0.1066*	*0.0223*	*0.2094*	8.74
GmSWEET46						
GmSWEET52	*123.3*	*395.7*	*0.1523*	*0.1129*	*0.7409*	12.48
GmSWEET19						
GmSWEET40	*202.3*	*529.7*	*0.1572*	*0.0604*	*0.3845*	12.88
GmSWEET28						
GmSWEET33	*212.7*	*522.3*	*0.089*	*0.0198*	*0.2224*	7.30
GmSWEET48						
GmSWEET9	*200*	*478*	*0.0352*	*0.0633*	*1.7948*	2.89
GmSWEET43						
GmSWEET32	*205.6*	*562.4*	*0.1257*	*0.0259*	*0.2057*	10.30
GmSWEET51						

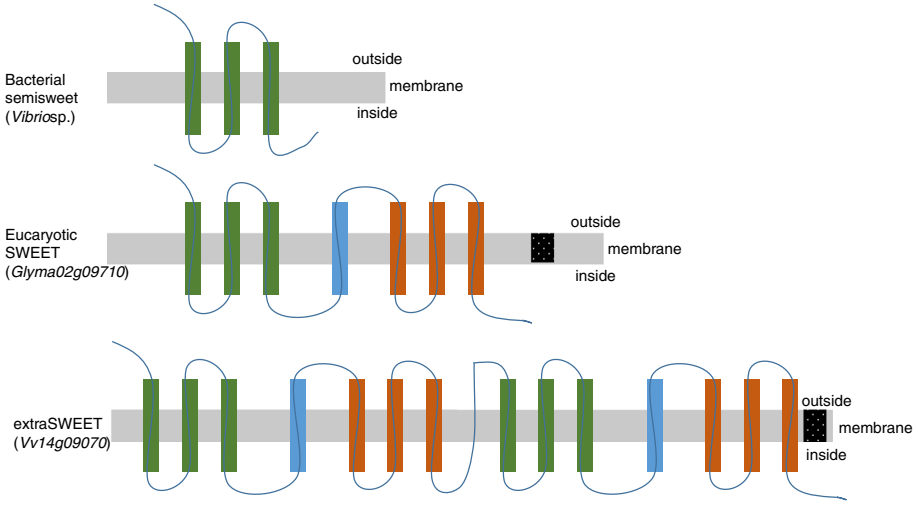

Fig. 4 Conserved domain architecture of SWEET proteins. SWEET proteins classified into 3 major types based on number of 3-TM domains (Additional file 4). Proteins with single 3-TM domain classified as semiSWEET [50]; proteins with two 3-TM (7 α-helical) classified as SWEET genes [43]; and proteins with four 3-TMs were named as extraSWEET genes. Black box shows the associated protein domains, position of associated protein domains could be -N or -C terminal (Additional file 4). Not drawn to scale

(Data not shown). In most SWEET genes, the second TM domain was found to be conserved rather than the first domain. A search for conserved domain architecture (using Conserved Domain Architecture Retrieval Tool [51]) resulted in three major types, as outlined in Fig. 4. These major types were further grouped into nine sub-types and they differed either in the position of *MtN3_slv* or they had regions with homology to other types of domains (e.g. receptor kinase, cuperdoxin, RNase H) and signal peptides (Fig. 4, Additional file 4).

As an interesting side finding, we found one SWEET protein from *V. vinifera* (*Vv14G09070*) that has duplication of 7-TM within the gene (Fig. 4, Additional file 4). This is a novel sub-type which we named extaSWEET. The extraSWEET gene could be another internal duplication of 7-TM, similar to the duplication of semi-SWEET (3-TM) to evolved in SWEET gene (7-TM) [43]. *V.vinifera* accumulates high levels of sugar compounds in their berries and this extraSWEET gene might have a role to mediate more sucrose transport. It has been reported that sucrose (*VvSUC*) and hexose (*VvHT*) transporter genes are preferentially expressed during berry development in *V. vinifera* [52]. In addition to the *VvSUC* and *VvHT*, it would be interesting to see the expression sites and function of *VvSWEET* (*Vv14G09070*) for long distance sugar transport during flower and/or berry development in *V. vinifera*.

The protein architecture and TM domains in soybean were conserved showing 36 SWEET genes with 7-TMs (SWEET), and the rest had less than 6 TMs (partial/semiSWEET) (Additional file 5). In addition to this, conserved *cis*-elements in the proximal promoter region (2 Kb upstream) among 52 *GmSWEET* genes were identified using INCLUSive MotifSampler [53]. Identification and comparing the *cis*-motif consensus pattern and discovery of expression modules within gene co-expression networks are crucial to understand the common regulatory networks. The top five significant *cis*-motif patterns were sampled from *GmSWEET* genes (Additional file 6). Motifs such as TBP binding sites, GT-2 (Grass TF 2), ATHB1 (*A. thaliana* Homeobox 1), HAHB4 (*H. annuus* Homeobox 4) and TaMYB80 (*T. aestivum* MYB80) were identified in SWEET gene promoters, indicating differential regulation and also they might have a putative role of sugar signaling [54] (Additional file 6). Interestingly, *cis*-motif elements of GT-2 and GT-3 were significantly enriched in soybean SWEET genes (Additional files 6 and 7). GT-2, −3 are plant transcriptional activators in higher plants and are involved in seed development and other diverse functions in rice, Arabidopsis and soybean [55, 56]. Further functional characterization of these *cis*-regulatory motifs and TFs (Transcription Factor) binding sites in *GmSWEET* genes will be helpful to understand the precise roles in development.

Soybean SWEET genes are highly expressed during reproduction and seed development

To understand the roles of specific *GmSWEET* genes in different developmental stages, we compared the expression profiles of all soybean SWEET genes using two publicly available RNA-seq datasets. The first dataset contains 14 tissues including whole seed at 11 stages of reproductive tissue development (flower, pod, and seeds) and three vegetative tissues (leaves, root, and nodules)

[57]. The second dataset contains 10 tissues including 6 reproductive tissues (floral buds, whole seeds at five stages of seed development i.e. globular, heart, cotyledon, early-maturation, dry), and four vegetative tissues (leaves, roots, stems, and seedlings) (GEO Accession GSE29163; Goldberg et. al. unpublished). Among all SWEET genes, *GmSWEET21* and *GmSWEET24* showed the highest expression in both of the datasets (Fig. 5a). The expression of 23 genes was either very low or undetectable in the datasets, hence they might be pseudo-genes or they might be expressed in certain tissues or conditions (Fig. 5a, Additional file 8). The gene expression pattern is varied in different developmental stages. Most of the genes were up-regulated during flower and seed development; several of them could be specific to these stages. It is noteworthy that the overall SWEET gene expression increased gradually during seed filling and then declined towards seed maturation (Fig. 5a, b). This suggests that the SWEET transporter plays a crucial role in nutrient unloading during seed development and seed filling. Overall results support earlier studies which concluded that most of the SWEET genes are related to reproductive development than other physiological processes [20, 58, 59].

In the present study 21 paralogous gene pairs for *GmSWEET* were identified (Fig. 3). The relationships between paralogous *GmSWEET* pairs with their expression pattern during development was compared. Nine out of 21 pairs showed a similar expression pattern and rest showed divergence in expression patterns (Additional file 9). For example, paralog pair *Glyma05g38340* (*GmSWEET11*) and *Glyma08g01310* (*GmSWEET21*) were up-regulated in cotyledonary tissue while simultaneously being down-regulated in leaf tissue. Similar expression levels of paralog genes suggests that they have retained the promoter element. Expression patterns of the remaining 12 paralog pairs has diverged (Fig. 5a), to either non-functionalization, neo-functionalization or sub-functionalization. Therefore, it would be interesting to see the expression pattern of those genes in soybean under different conditions. In soybean, SWEET genes are also associated with the iron deficiency [60]. Lauter

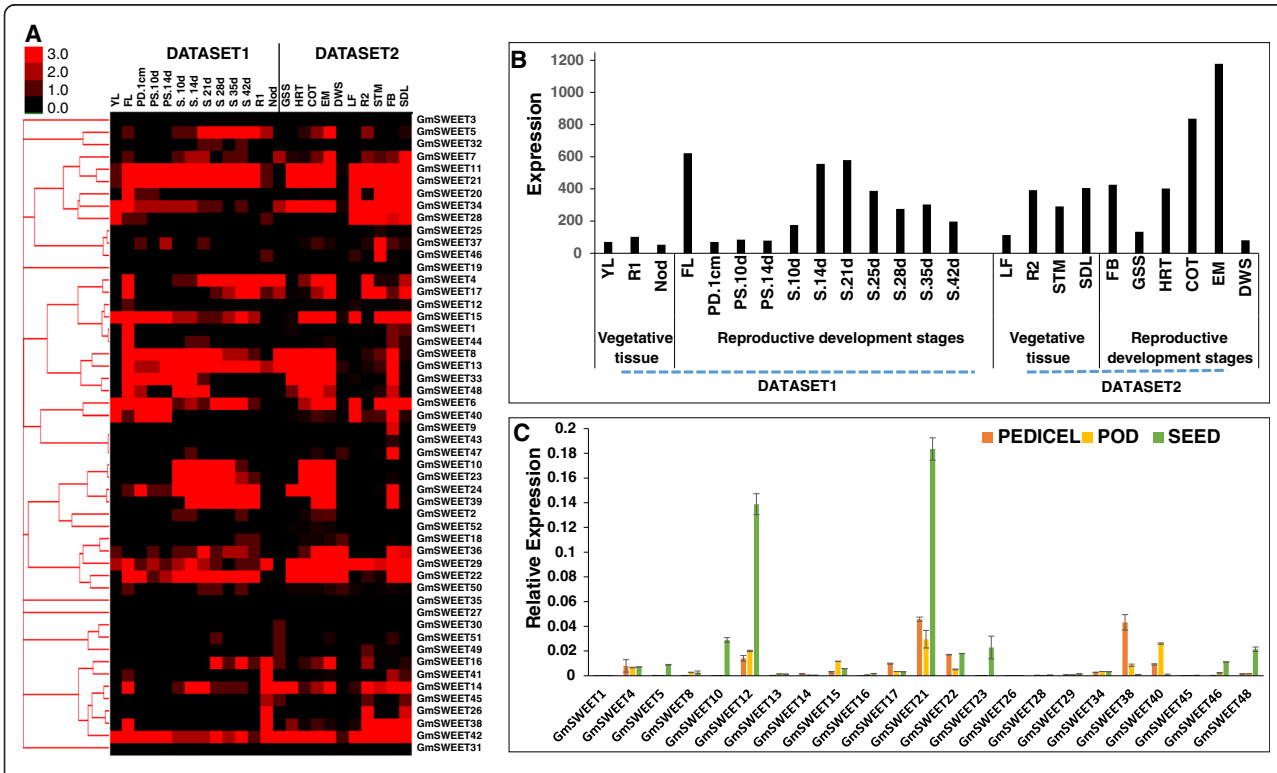

Fig. 5 Expression profiles of soybean SWEET genes in different tissues. **a** Hierarchical cluster of expression profiles from two RNA-seq datasets in 24 tissues covering the whole life cycle of soybean (Williams 82). Sources of the samples are as follows: Dataset 1 - YL (young leaves), FL (flower), PD.1 cm (one cm pod), PS.10d (pod shell 10 Days After Flowering (DAF)), PS.14d (pod shell 14DAF), S.10d (seed 10DAF), S.14d (seed 14 DAF), S.21d (seed 21DAF), S.25d (seed 25DAF), S.28d (seed 28DAF), S.35d (seed 35DAF), S.42d (seed 42DAF), R1 (root), and Nod (nodule); Dataset 2 - GSS (globular stage whole seed), HRT (heart stage), COT (cotyledonary stage), EM (early maturation), DWS (dry whole seed), LF (leaf), R2 (root), STM (stem), FB (floral bud), and SDL (seedling). **b** Expression pattern of 52 *GmSWEET* genes. Bars show the expression of all genes in different developmental stages from both datasets. For simplicity, the datasets were marked with vegetative and reproductive stages. **c** qRT-PCR analysis of 21 selected *GmSWEET* genes in pedicel, pod, and seed tissues

et. al. (2014) observed the repression of two SWEET genes (*Glyma05g38340* and *Glyma08g01310*) and other sucrose transporter genes in the leaves one hour after iron stress and concluded that SWEET genes might play a role in regulation of the SnRK1/TOR (SNORKEL) signaling pathway in response to iron deficiency [60].

Examination of SWEET gene expression in reproductive tissue by qRT-PCR

To confirm the expression patterns determined by the RNA-seq analysis, qRT-PCR was employed to analyze the expression patterns of 23 genes in three reproductive development tissues of soybean, Williams 82 (W82), specifically pedicel, pods, and developing seeds (Fig. 5c). The expression patterns (Fig. 5c) were largely consistent with those obtained by the RNAseq analysis (Fig. 5a), even though some smaller variations can be seen. GmSWEET 12 and 21 were highly expressed in all three developmental stages, but in the seeds they are so abundant that the total relative SWEET gene expression far exceeds that of the other tissues (Fig. 5c). The expression of GmSWEET genes 5, 10, 23, and 48 were also much higher in seeds than in the other tissues, and may be considered seed-specific. In pods, GmSWEET 12, 21, and 40 had comparatively higher expression, and in pedicels the expression of GmSWEET 12, 21, and 38 stands out.

Exploring natural variation in *GmSWEET* genes using soybean whole genome re-sequencing data

The elucidation of the soybean SWEET genes gave us an unprecedented opportunity to obtain a comprehensive overview of the allelic variation in soybean whole genome re-sequencing data. The wealth of whole genome resources of soybean provides a unique angle to study natural variation in germplasm and further allows functional characterization of the particular gene [61–63]. Complete genome sequences for 106 soybean genotype, sequenced at approximately 15X coverage, were obtained from the Soybean Genetics and Genomics Laboratory at The University of Missouri (Valliyodan et. al. Unpublished) and analyzed for synonymous and non-synonymous SNPs, premature stop codon and haplotype variation in selected *GmSWEET* genes. In Arabidopsis, *AtSWEET11* and *–12* double mutants accumulated sucrose in the leaves and had lower levels in the phloem, identifying them as the long sought main sucrose effluxers in the leaf sugar export pathway [26]. It has been observed that when *AtSWEET17* expression is reduced, either by induced or natural variation, fructose accumulates in the leaves, suggesting an enhanced storage capacity [64]. Site directed mutagenesis of *AtSWEET1* at four conserved positions (P23T, Y57A, G58D, and G180D) led to abolishment of glucose transport activity in a yeast complementation assay. Also, SNP in the coding or promoter region can also abolish protein localization and function [43]. In the present study, wide natural variations were observed in non-synonymous SNPs and a total of 37 SNPs were observed in 21 (~40 %) *GmSWEET* genes (Table S5). *GmSWEET41* (*Glyma15g27530*) showed a premature stop codon in the 1st exon in 15 sequenced lines.

To understand and visualize the genetic variation in whole genome re-sequencing data for the SWEET genes, a cluster of genes (*GmSWEET15, 16, and 17*) including their 2 kb promoter region was examined. The haplogroup gave three major distinct clusters based on the SNP variation in promoter and coding regions similar or dissimilar to the soybean reference genome, W82 (Additional file 10). As sugar derivatives are associated with SWEET genes [8, 43], we further examined the association between the haplogroup cluster and different sugar content (sucrose, raffinose, and stachyose) in soybean seeds and observed a correlation between three SNP-haplogroups and average sucrose content. The SNP-haplogroup similar to reference genotype W82 showed intermediate sucrose concentration of average 5.26 ± 0.14 %. The other two groups were distinct from W82 haplogroup showing an average sucrose concentration of 4.8 ± 0.4 % and 5.5 ± 0.28 %, (Additional file 10). Out of 10 wild soybean lines (*G. soja*), seven lines were identified in the first haplogroup which showed a relatively lower sucrose content. No significant association was found for raffinose and stachyose concentrations. It has been reported that the transport of Raffinose family oligosaccharides (RFOs) are not detectable when associated with apoplastic loading [23, 65] and several higher plants accumulate RFO during the seed maturation process [66], hence SWEET genes might have no role in efflux for RFOs. However, to fully understand their roles, detailed functional characterization of the individual gene is needed.

Discussion and conclusions

In-silico analysis and phylogenetic studies generate valuable information on the evolutionary and functional relationships between genes of different species, genomic complexity, and lineage-specific adaptations. Previous work on sugar transporter genes SWEET *(MtN3_slv)*, along with the rapidly expanding availability of genomics sequence data has enabled us to examine the SWEET content of multiple plant genomes.

The SWEET gene family has been studied in Arabidopsis [15, 26], rice [20, 58, 59] and bacteria [43, 50]. However, this family has not previously been studied in soybean. Here, we explore these genes in soybean with an analysis of their phylogeny, gene structure, domain architecture, expression profiles and natural genetic variation. A total of 52 full-length SWEET genes were identified in the soybean genome, which is highest among the analyzed

plants and implies a genome expansion. The exon/intron layouts and the TM motifs were quite conserved when compared to the paralogs. A phylogenetic tree was constructed (Fig. 2) to identify putative orthologous and paralogous SWEET genes and to study the pattern of the SWEET gene family expansion in the course of evolution.

The salt water living chlorophyta algae *O. tauri, O. lucimarius and Micromonos sp*. have only a single gene. On the other hand, the fresh water algae, *V. carteri and C. reinhardtii,* contain 2 and 3 SWEET genes, respectively. This leads us to suspect that during the transition phase to fresh water, a more involved mechanism for sugar transport was required by environmental conditions. The evolution to multi-cellularity led to further expansion of the SWEET gene family. Recent studies on the evolution of the SUT transporter family showed that divergence of different SUT types were likely associated with evolution of vascular cambium and phloem transport [34]. Higher plants evolved phloem for long-distance, source-to-sink transport. Although different phloem loading strategies are recognized, lineages that evolved apoplasmic phloem loading required a mechanism for efflux from phloem parenchyma and subsequent energized uptake into the companion cell/sieve element complex, SWEETs provided the former function [6]. *P. patens* is an early diverging land plant and many families of *P. patens* genes for metabolic enzymes (e.g. cytokinin [67], glutathione [68], pectin [69]) have large copy numbers. *P. patens* has only a primitive protophloem, and the increase in the SWEET genes here could be due to the recent genome duplication [70], without the new genes necessarily having acquired differentiated functionalities. *S. moellendorffii* does have a phloem, and the number of SWEET genes here approaches that of many angiosperms (Fig. 1).

The expansion of a gene family in higher plants indicates the differentiation of physiological function of each isoform in terms of the expression site and the regulatory manner which subsequently helps the organism to adapt in different environmental conditions. The internal duplication of the 3-TM (semiSWEET) gene must have happened early to give rise to new genes with 7-TMs (SWEET) which allow a more sophisticated sucrose transport [43, 50, 71]. Here we also report a novel gene in *V. vinifera* which has further duplicated the TM regions. Collectively, phylogenetic and domain studies imply that biological, physiological or environmental conditions forces particular gene families to evolve and expand. As evolution of the higher plants have progressed, some species have acquired further SWEET genes (Fig. 1). This suggests that sugar transport evolution has followed as new plant structures and adaptations to new ecological niches have arisen.

The SWEET genes play a diverse functional role during plant development which is evident from their expression patterns in other plant species [25, 40, 43, 50, 58] and soybean (this report). In rice and Arabidopsis, the expression of the SWEET genes were relatively higher in flower, pollen, embryo sac and seeds suggesting their roles in reproductive tissue development [15, 19, 20]. In rice two members of the SWEET gene family were highly expressed in panicles and anthers and were associated with fertility and seed size [20, 58, 59]. In Arabidopsis, *AtSWEET8* was expressed in the embryo sac suggesting that it might regulate female gametocyte development [72]. Developing seeds are the strongest sink tissues in many plants and they need a higher carbon source for development which implies that nutrient transporters including the SWEET genes might be key component for their development. In Arabidopsis, *AtSWEET11* and *−12* showed a higher expression in leaves and had important roles in leaf sucrose export [15]. The comparison of *AtSWEET11* and *−12* expression pattern with soybean orthologs *GmSWEET6,* and *−15* showed a relatively higher expression in leaves, suggesting that these genes also might have similar role in leaf sucrose export.

Yuan and Wang [20] and Chen [25] have reviewed the functional role of SWEET genes in different tissues, pathogen infestation, and environmental responses. Interestingly, *GmSWEET13*, *14* and *15* fall under the fungal disease resistance QTL on chromosome 6 in soybean [73]. It has been proven that fungal and bacterial symbionts induce SWEET gene expression for nutritional gain during pathogen infestation [15, 25, 40, 74, 75]. The statement that most of the reported SWEET genes are associated with reproductive development tissue is corroborated in this study using soybean transcriptome datasets. The transciptome and qRT-PCR data showed that multiple SWEET genes are expressed at higher levels in tissues involved in reproductive development. Relatively higher expression of *GmSWEET5*, *−10*, *−23* and *−48* in the seed tissue, suggest that collectively these genes might assist the movement of sucrose in the developing soybean seeds. Unloading of nutrient in the developing seeds occurs from the seed coat [32, 76]. In the developing legume seeds (*P. vulgaris* and *P. sativum*), a suite of sucrose transporters are expressed at a higher levels in seed coat tissue to facilitate the movement of sucrose [36]. Sugar availability, starch content, and cytokinin levels are involved in the regulation of abscission of soybean flowers, the delay of which hampers seed development and leads to yield loss in soybean [77–79]. Soybean flower abortion is primarily caused by deficiency in or competition for photo-assimilates and nutrients among growing organs.

Beside the expression level, the genetic variation (natural or induced) also enforces the functionality of SWEET genes and causes a variation in phenotype [43, 64]. Mutation in the SWEET gene or abolishing the activation of the

SWEET promoter leads to resistance to bacterial pathogens in rice [5, 59, 80]. Identification of several non-synonymous SNPs and large effect SNPs in *GmSWEETs* are expected to affect the integrity of encoded proteins. Additionally, exploring SNP-haplotype diversity using whole-genome sequencing data mining provides a powerful resource for investigating diversity in a particular gene family [81–83]. The data presented here, using a cluster of genes on chromosome 6 (*GmSWEET13, −14* and *−15*), showed the association between the SNP-haplogroups and sucrose content in seeds. The allelic variation data presented in this study provides a valuable resource for association studies between the SNPs and important agronomic traits, although intensive studies with each candidate gene are required to examine this inference. Overall, the SWEET gene family signifies its role as a key component in reproductive tissue development, nutrient unloading and pathogen resistance. Manipulating SWEET expression in specific tissues (phloem sap, pedicel, and developing seeds) could enhance sugar delivery to developing seeds to increase yield.

Methods

Sequence and database search for SWEET gene family

SWEET (*MtN3_slv)* gene families were identified from 25 completely sequenced genomes representing the plant lineage (green plants) including members from unicellular green algae to multicellular plants (Fig. 1, Additional file 1). The protein BLAST search was performed using *AtSWEET11* as a query sequence in Plaza [42] (http://bioinformatics.psb.ugent.be/plaza/news/index) and Phytozome [84] (http://www.phytozome.org) databases and the sequences were retrieved from the corresponding plant genome annotation resources and analyzed. The multiple sequence alignment was performed using MUSCLE program [85] and partial and redundant sequences were excluded. All proteins were examined for presence of *MtN3_slv* related TM domains (IPR018179) using Interpro database [86] (http://www.ebi.ac.uk/). *Glycine max* SWEET genes were designated as *GmSWEET1* to *GmSWEET52*.

Phylogenetic analysis

To understand the phylogenetic relationship, 173 SWEET genes from 13 species representing major clades were analyzed. Protein sequences were analyzed by the neighbor-joining (NJ) method [87] with genetic distance calculated by MEGA5.1 [88] (www.megasoftware.net/). The numbers at the nodes represent bootstrap percentage value based on 1,000 replications.

Identification of conserved domains and *cis*-motif pattern

The Conserved Domain Architecture Retrieval Tool (CDART) [51] (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi) was searched using Arabidopsis *AtSWEET11* as a query protein sequence (Additional file 4). Several *MtN3_slv* TM domains were preceded which were grouped into three major architecture based on 3-TMs and associated proteins (Fig. 4). Identification of the exon/intron organization of SWEET genes was performed by aligning cDNAs with their corresponding genomic DNA sequences and were also obtained by using the Plaza comparative database. *Cis* regulatory elements were identified by searching 2 kb upstream of the 5' translation start base for all of the soybean SWEET genes using INCLUSive MotifSampler [89]. 2 kb upstream sequences were annotated by similarity search (p value <0.05, motif score >5) with known plant transcription binding sites and motifs available in the Athamap database [90] (www.athamap.de, Additional files 6 and 7).

Soybean SWEET gene chromosomal location and gene duplication

The location of soybean SWEET genes was determined based on their physical positions on chromosomes corresponding to their locus numbers in the SoyKB browser [91]. The duplication of SWEET genes on segmentally duplicated regions was determined using Plaza 2.5 whole genome mapping tool (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v2_5/genome_mapping/genome_mapping), and were visualized using genome search and synteny view tool (CViT) (http://comparative-legumes.org/) [92]. The comparative duplicate block representing homologous chromosome segments were anchored on 15 out of 20 soybean chromosomes and indicated by tandem/block duplication (Fig. 3).

Calculation of Ka/Ks values

Non-synonymous (Ka) to synonymous (Ks) substitution rates were used to estimate the selection mode for all orthologous gene pairs of soybean SWEET family [48]. Subsequently, the PAL2NAL program (http://www.bork.embl.de/pal2nal/) was used to convert a multiple sequence alignment of proteins and the corresponding DNA (or mRNA) sequences into a codon alignment [93]. PAL2NAL automatically calculates Ks and Ka by the CODEML program in PAML. The divergence time (T) was calculated by $T = Ks/(2 \times 6.1 \times 10^{-9}) \times 10^{-6}$ MYA, where 6.1×10^{-9} is divergence rate in millions of years translated from Ks value [49].

RNA-seq datasets and qRT-PCR analysis

Genome-wide public RNA-seq datasets (Reads/Kb/Million (RPKM) normalized data) for soybean developmental stages were downloaded from soybean RNA-seq Atlas [57] and Gene Expression Omnibus (GEO) database (accession number GSE29163) from Goldberg et. al.

(Unpublished). Sources of the samples for first dataset are as follows: YL (young leaves), FL (flower), PD.1 cm (one cm pod), PS.10d (pod shell 10 Days After Flowering (DAF)), PS.14d (pod shell 14DAF), S.10d (seed 10DAF), S.14d (seed 14 DAF), S.21d (seed 21DAF), S.25d (seed 25DAF), S.28d (seed 28DAF), S.35d (seed 35DAF), S.42d (seed 42DAF), R1 (root), and Nod (nodule). Sources of the samples for second dataset are as follows: GSS (Globular stage whole seed), HRT (Heart stage), COT (Cotyledonary stage), EM (Early maturation), DWS (Dry whole seed), LF (Leaf), R2 (Root), STM (Stem), FB (Floral bud), and SDL (Seedling). Average linkage method provided in Cluster 3.0 was used to cluster gene and tissue types and visualized using TreeView software [94].

Total RNA was extracted from soybean pedicel, pod, and seed tissues using a Qiagen RNeasy mini kit (Qiagen, CA, USA). First strand cDNA from 1 μg of total RNA was synthesized by using Superscript III reverse transcriptase (Invitrogen) with oligo(dT) primer. Primers for quantitative reverse transcription PCR (qPCR) were designed using Primer3 (http://frodo.wi.mit.edu) (Additional file 11). Quantitative RT-PCR was performed using cDNA product in a 10 μl reaction volume using Maxima SYBR Green/ROX qPCR master mix (Thermo, USA) on ABI7900HT detection system (Life Technologies, NY, USA). Three biological replicates and two technical replicates were used for analysis. The PCR conditions were: 50 °C for 2 min., 95 °C for 10 min., then 40 cycles of 95 °C for 15 sec., and 60 °C for 1 min. To normalize the gene expression, Actin (*Glyma18g52780*) was used as an internal control.

Analysis of sequence variants, non-synonymous SNP and haplotype variation

One hundred and six soybean lines with carbohydrate phenotypes (sucrose, stachyose, and raffinose) and whole genome re-sequencing (sequencing depth approximately 15X) data were obtained for soybean SWEET genes from Soybean Genetics and Genomics Laboratory at the University of Missouri (Valliyodan et. al. Unpublished). The processed data was aligned to the Williams 82 Gmax v9.0 from Phytozome as the reference genome [46]. SNPs were identified using an in-house built pipeline using with SOAP3 [95] and were analyzed for possible synonymous/non-synonymous SNP variation annotations using SnpEFF [96] and v9.0 gene models from Phytozome (Additional file 12). SNP haplotypes were examined by generating map and genotype data files using TASSEL 5.0 program [97] and then clustering pictorial output for a specific genic region was visualized using FLAPJACK software [98].

Availability of supporting data

All supporting data of this article are included as additional files.

Additional files

Additional file 1: SWEET gene family across 25 plant genomes.

Additional file 2: Amino acid sequences of 173 SWEETs from 13 species.

Additional file 3: Gene organization of soybean SWEET orthologs genes.

Additional file 4: Conserved domain and associated protein architecture in SWEET (*MtN3_slv*) gene family.

Additional file 5: Soybean SWEET protein Transmembrane Helix prediction obtained from TMHMM 2.0 tool [99].

Additional file 6: *cis*-motif analysis of soybean SWEET genes. Conserved motifs identified in proximal promoter region of SWEET gene family using INCLUSive MotifSampler and its similarity with known motifs available in Athmap database. Similarity search performed using STAM tool (www.benoslab.pitt.edu/stamp).

Additional file 7: Detailed information of *cis*-motifs in upstream promoter region of soybean SWEET genes.

Additional file 8: *GmSWEET* gene transcriptome profile in public datasets.

Additional file 9: Expression profile of *GmSWEET* paralogue gene pairs.

Additional file 10: SNP-Haplotype analysis of SWEET gene cluster on chromosome 6. Hierarchical clustering showed the association between average sucrose content and SNPs haplogroups in 106 lines. Base position identical to reference (Williams 82) are light sky blue, black – different, gray-missing data. Blue colored bar on top showing approximate position of *GmSWEET15, −16* and *−17*.

Additional file 11: Primer sequences of 23 selected *GmSWEET* genes for RT-qPCR analysis.

Additional file 12: Identification of non-synonymous SNP in 52 *GmSWEET* from 106 soybean re-sequencing data (Valliyodan et al. Unpublished). The Seq_id with underline represents wild soybean lines (*G. soja*).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GP, BV and HTN conceived and planned the study. GP, SP, BN and HTN contributed to drafting the manuscript. GP, RD and BN performed data mining, analysis and interpretation. RD, LL, YL, TJ and DX carried out *cis*-element analysis. BV provided whole genome re-sequencing data and GP and HS performed analysis. MZ, LS and JC assisted with RNAseq and qRT-PCR analysis. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge Missouri Soybean Merchandising Council for the funding support (Grant # 288 and # 368). The authors also would like to thank Theresa A. Musket for reviewing and editing the manuscript.

Author details

[1]National Center for Soybean Biotechnology and Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA. [2]Department of Plant Biology and Forest Genetics and Linnean Center for Plant Biology, Swedish University of Agricultural Sciences, Uppsala, Sweden. [3]Department of Computer Science, Informatics Institute, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA. [4]Current address: Agronomy College of Shenyang Agricultural University, Shenyang, China.

Received: 29 December 2014 Accepted: 26 June 2015
Published online: 11 July 2015

References

1. Rolland F, Moore B, Sheen J. Sugar sensing and signaling in plants. Plant Cell. 2002;14(Supplement):S185–205.
2. Wind J, Smeekens S, Hanson J. Sucrose: metabolite and signaling molecule. Phytochem. 2010;71(14):1610–4.
3. Ayre BG. Membrane-transport systems for sucrose in relation to whole-plant carbon partitioning. Mol Plant. 2011;4:ssr014.
4. Sauer N. Molecular physiology of higher plant sucrose transporters. FEBS Lett. 2007;581(12):2309–17.
5. Ruan Y-L. Sucrose metabolism: gateway to diverse carbon use and sugar signaling. Annu Rev Plant Biol. 2014;65:33–67.
6. Baker RF, Leach KA, Braun DM. SWEET as sugar: new sucrose effluxers in plants. Mol Plant. 2012;5(4):766–8.
7. Turgeon R, Wolf S. Phloem transport: cellular pathways and molecular trafficking. Annu Rev Plant Biol. 2009;60:207–21.
8. Lemoine R, La Camera S, Atanassova R, Dedaldechamp F, Allario T, Pourtau N, et al. Source-to-sink transport of sugar and regulation by environmental factors. Front Plant Sci. 2013;4:272.
9. Braun DM, Slewinski TL. Genetic control of carbon partitioning in grasses: roles of sucrose transporters and tie-dyed loci in phloem loading. Plant Physiol. 2009;149(1):71–81.
10. Rennie EA, Turgeon R. A comprehensive picture of phloem loading strategies. Proc Natl Acad Sci U S A. 2009;106(33):14162–7.
11. Lohaus G, Burba M, Heldt H. Comparison of the contents of sucrose and amino acids in the leaves, phloem sap and taproots of high and low sugar-producing hybrids of sugar beet (*Beta vulgaris* L.). J Exp Bot. 1994;45(8):1097–101.
12. Slewinski TL, Meeley R, Braun DM. Sucrose transporter1 functions in phloem loading in maize leaves. J Exp Bot. 2009;60(3):881–92.
13. Srivastava AC, Ganesan S, Ismail IO, Ayre BG. Functional characterization of the Arabidopsis *AtSUC2* sucrose/H+ symporter by tissue-specific complementation reveals an essential role in phloem loading but not in long-distance transport. Plant Physiol. 2008;148(1):200–11.
14. Aoki N, Hirose T, Scofield GN, Whitfeld PR, Furbank RT. The sucrose transporter gene family in rice. Plant Cell Physiol. 2003;44(3):223–32.
15. Chen L-Q, Hou B-H, Lalonde S, Takanaga H, Hartung ML, Qu X-Q, et al. Sugar transporters for intercellular exchange and nutrition of pathogens. Nature. 2010;468(7323):527–32.
16. Sonnewald U. SWEETS–the missing sugar efflux carriers. Front Plant Sci. 2011;2:1–2.
17. Kühn C. A comparison of the sucrose transporter systems of different plant species. Plant Biol. 2003;5(3):215–32.
18. Contim LAS, Waclawovsky AJ, Delú-Filho N, Pirovani CP, Clarindo WR, Loureiro ME, et al. The soybean sucrose binding protein gene family: genomic organization, gene copy number and tissue-specific expression of the SBP2 promoter. J Exp Bot. 2003;54(393):2643–53.
19. Guan Y-F, Huang X-Y, Zhu J, Gao J-F, Zhang H-X, Yang Z-N. RUPTURED POLLEN GRAIN1, a member of the *MtN3/saliva* gene family, is crucial for exine pattern formation and cell integrity of microspores in Arabidopsis. Plant Physiol. 2008;147(2):852–63.
20. Yuan M, Wang S. Rice *MtN3/saliva* family genes and their homologues in cellular organisms. Mol Plant. 2013;6:sst035.
21. Gamas P, de Carvalho NF, Lescure N, Cullimore JV. Use of a subtractive hybridization approach to identify new Medicago truncatula genes induced during root nodule development. MPMI. 1996;9(4):233–42.
22. Artero RD, Terol-Alcayde J, Paricio N, Ring J, Bargues M, Torres A, et al. Saliva, a new *Drosophila* gene expressed in the embryonic salivary glands with homologues in plants and vertebrates. Mech Dev. 1998;75(1):159–62.
23. Slewinski TL. Diverse functional roles of monosaccharide transporters and their homologs in vascular plants: a physiological perspective. Mol Plant. 2011;4(4):641–62.
24. Braun DM. SWEET! The pathway is complete. Science. 2012;335(6065):173–4.
25. Chen LQ. SWEET sugar transporters for phloem transport and pathogen nutrition. New Phytol. 2014;201(4):1150–5.
26. Chen L-Q, Qu X-Q, Hou B-H, Sosso D, Osorio S, Fernie AR, et al. Sucrose efflux mediated by SWEET proteins as a key step for phloem transport. Science. 2012;335(6065):207–11.
27. Streubel J, Pesce C, Hutin M, Koebnik R, Boch J, Szurek B. Five phylogenetically close rice SWEET genes confer TAL effector-mediated susceptibility to *Xanthomonas oryzae* pv. oryzae. New Phytol. 2013;200(3):808–19.
28. Denancé N, Szurek B, Noël LD. Emerging functions of nodulin-like proteins in non-nodulating plant species. Plant Cell Physiol. 2014;55(3):469–74.
29. Doidy J, Grace E, Kühn C, Simon-Plas F, Casieri L, Wipf D. Sugar transporters in plants and in their interactions with fungi. Trends Plant Sci. 2012;17(7):413–22.
30. Patrick JW. PHLOEM UNLOADING: sieve element unloading and post-sieve element transport. Annu Rev Plant Physiol Plant Mol Biol. 1997;48(1):191–222.
31. Baud S, Dubreucq B, Miquel M, Rochat C, Lepiniec L. Storage reserve accumulation in Arabidopsis: metabolic and developmental control of seed filling. Am Soc Plant Biologists. 2008;6:e0113.
32. Weber H, Borisjuk L, Wobus U. Molecular physiology of legume seed development. Annu Rev Plant Biol. 2005;56:253–79.
33. Zhang W-H, Zhou Y, Dibley KE, Tyerman SD, Furbank RT, Patrick JW. Review: Nutrient loading of developing seeds. Funct Plant Biol. 2007;34(4):314–31.
34. Lalonde S, Tegeder M, Throne-Holst M, Frommer W, Patrick J. Phloem loading and unloading of sugars and amino acids. Plant Cell Environ. 2003;26(1):37–56.
35. Marschner H, Marschner P. Marschner's mineral nutrition of higher plants. London: Academic press; 2012.
36. Zhou Y, Qu H, Dibley KE, Offler CE, Patrick JW. A suite of sucrose transporters expressed in coats of developing legume seeds includes novel pH-independent facilitators. Plant J. 2007;49(4):750–64.
37. Ludewig F, Flügge U-I. Role of metabolite transporters in source-sink carbon allocation. Front Plant Sci. 2013;4:231.
38. Weschke W, Panitz R, Gubatz S, Wang Q, Radchuk R, Weber H, et al. The role of invertases and hexose transporters in controlling sugar ratios in maternal and filial tissues of barley caryopses during early development. Plant J. 2003;33(2):395–411.
39. Wei X, Liu F, Chen C, Ma F, Li M. The *Malus domestica* sugar transporter gene family: identifications based on genome and expression profiling related to the accumulation of fruit sugars. Front Plant Sci. 2014;5:569.
40. Yang B, Sugio A, White FF. *Os8N3* is a host disease-susceptibility gene for bacterial blight of rice. Proc Natl Acad Sci U S A. 2006;103(27):10503–8.
41. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2011;40:gkr1065.
42. Van Bel M, Proost S, Wischnitzki E, Movahedi S, Scheerlinck C, Van De Peer Y, et al. Dissecting plant genomes with the PLAZA comparative genomics platform. Plant Physiol. 2011;111:189514.
43. Xuan YH, Hu YB, Chen L-Q, Sosso D, Ducat DC, Hou B-H, et al. Functional role of oligomerization for bacterial and plant SWEET sugar transporter family. Proc Natl Acad Sci U S A. 2013;110(39):E3685–94.
44. Severin AJ, Cannon SB, Graham MM, Grant D, Shoemaker RC. Changes in twelve homoeologous genomic regions in soybean following three rounds of polyploidy. Plant Cell. 2011;23(9):3129–36.
45. Soltis DE, Visger CJ, Soltis PS. The polyploidy revolution then… and now: Stebbins revisited. Am J Bot. 2014;101(7):1057–78.
46. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178–83.
47. Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, May G, et al. The fate of duplicated genes in a polyploid plant genome. Plant J. 2013;73(1):143–53.
48. Li W-H, Gojobori T, Nei M. Pseudogenes as a paradigm of neutral evolution. Nature. 1981;292(5820):237–9.
49. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. Science. 2000;290(5494):1151–5.
50. Xu Y, Tao Y, Cheung LS, Fan C, Chen L-Q, Xu S, et al. Structures of bacterial homologues of SWEET transporters in two distinct conformations. Nature. 2014;515:448–52.
51. Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. Genome Res. 2002;12(10):1619–23.
52. Afoufa-Bastien D, Medici A, Jeauffre J, Coutos-Thevenot P, Lemoine R, Atanassova R, et al. The *Vitis vinifera* sugar transporter gene family: phylogenetic overview and macroarray expression profiling. BMC Plant Biol. 2010;10(1):245.
53. Thijs G, Moreau Y, De Smet F, Mathys J, Lescot M, Rombauts S, et al. INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. Bioinformatics. 2002;18(2):331–2.
54. Klepek YS, Volke M, Konrad KR, Wippel K, Hoth S, Hedrich R, et al. *Arabidopsis thaliana* POLYOL/MONOSACCHARIDE TRANSPORTERS 1 and 2: fructose and xylitol/H+ symporters in pollen and young xylem cells. J Exp Bot. 2009;61:erp322.
55. Smalle J, Kurepa J, Haegman M, Gielen J, Van Montagu M, Van Der Straeten D. The trihelix DNA-binding motif in higher plants is not restricted to the transcription factors GT-1 and GT-2. Proc Natl Acad Sci U S A. 1998;95(6):3318–22.

56. Zhou D-X. Regulatory mechanism of plant gene transcription by GT-elements and GT-factors. Trends Plant Sci. 1999;4(6):210–4.
57. Severin AJ, Woody JL, Bolon Y-T, Joseph B, Diers BW, Farmer AD, et al. RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol. 2010;10(1):160.
58. Antony G, Zhou J, Huang S, Li T, Liu B, White F, et al. Rice xa13 recessive resistance to bacterial blight is defeated by induction of the disease susceptibility gene *Os-11 N3*. Plant Cell. 2010;22(11):3864–76.
59. Chu Z, Yuan M, Yao J, Ge X, Yuan B, Xu C, et al. Promoter mutations of an essential gene for pollen development result in disease resistance in rice. Genes Dev. 2006;20(10):1250–5.
60. Lauter ANM, Peiffer GA, Yin T, Whitham SA, Cook D, Shoemaker RC, et al. Identification of candidate genes involved in early iron deficiency chlorosis signaling in soybean (*Glycine max*) roots and leaves. BMC Genomics. 2014;15(1):702.
61. Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010;42(12):1053–9.
62. Qi X, Li M-W, Xie M, Liu X, Ni M, Shao G, et al. Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Comm. 2014;5:4340.
63. Patil G. Identification of sequence variants in candidate genes for Oil content using whole genome Re-sequencing of soybean germplasm. In: Plant and animal genome XXII conference. San Diego, CA Plant and Animal Genome; 2014.
64. Chardon F, Bedu M, Calenge F, Klemens PA, Spinner L, Clement G, et al. Leaf fructose content is controlled by the vacuolar transporter SWEET17 in *Arabidopsis*. Curr Biol. 2013;23(8):697–702.
65. Zimmermann MH, Ziegler H. List of sugars and sugar alcohols in sieve-tube exudates. New Ser: Encycl Plant Physiol; 1975.
66. Ayre BG, Keller F, Turgeon R. Symplastic continuity between companion cells and the translocation stream: long-distance transport is controlled by retention and retrieval mechanisms in the phloem. Plant Physiol. 2003;131(4):1518–28.
67. Patil G, Nicander B. Identification of two additional members of the tRNA isopentenyltransferase family in Physcomitrella patens. Plant Mol Biol. 2013;82(4–5):417–26.
68. Liu Y-J, Han X-M, Ren L-L, Yang H-L, Zeng Q-Y. Functional divergence of the glutathione S-transferase supergene family in Physcomitrella patens reveals complex patterns of large gene family evolution in land plants. Plant Physiol. 2013;161(2):773–86.
69. McCarthy TW, Der JP, Honaas LA, Anderson CT. Phylogenetic analysis of pectin-related gene families in Physcomitrella patens and nine other plant species yields evolutionary insights into cell walls. BMC Plant Biol. 2014;14(1):79.
70. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, et al. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science. 2008;319(5859):64–9.
71. Keller R, Ziegler C, Schneider D. When two turn into one: evolution of membrane transporters from half modules. Biol Chem. 2014;395(12):1379–88.
72. Talbot NJ. Cell biology: Raiding the sweet shop. Nature. 2010;468(7323):510–1.
73. Wang J-L, Liu C-Y, Wang J, Qi Z-M, Li H, Hu G-H, et al. An integrated QTL Map of fungal disease resistance in soybean (*glycine max* L. Merr): a method of meta-analysis for mining R genes. Agric Sci China. 2010;9(2):223–32.
74. Yuan M, Chu Z, Li X, Xu C, Wang S. The bacterial pathogen Xanthomonas oryzae overcomes rice defenses by regulating host copper redistribution. Plant Cell. 2010;22(9):3164–76.
75. van Ooij C. Pathogenesis: The SWEET life of pathogens. Nat Rev Microb. 2011;9(1):4–5.
76. Patrick J, Offler C. Post-sieve element transport of sucrose in developing seeds. Funct Plant Biol. 1995;22(4):681–702.
77. Antos M, Wiebold W. Abscission, total soluble sugars, and starch profiles within a soybean canopy. Agron J. 1984;76(5):715–9.
78. Nagel L, Brewster R, Riedell W, Reese R. Cytokinin regulation of flower and pod set in soybeans (*Glycine max* (L.) Merr.). Ann Bot. 2001;88(1):27–31.
79. Dybing CD, Reese ZN. Nitrogen and carbohydrate nutrient concentrations and flower Set in soybean *glycine max* (L.) merr.). J Biol Sci. 2008;8(1):24–33.
80. Li C, Wei J, Lin Y, Chen H. Gene silencing using the recessive rice bacterial blight resistance gene xa13 as a new paradigm in plant breeding. Plant Cell Rep. 2012;31(5):851–62.
81. Fernandez L, Le Cunff L, Tello J, Lacombe T, Boursiquot JM, Fournier Level A, et al. This P: Haplotype diversity of *VvTFL1A* gene and association with cluster traits in grapevine (*V. vinifera*). BMC Plant Biol. 2014;14(1):209.
82. Langewisch T, Zhang H, Vincent R, Joshi T, Xu D, Bilyeu K. Major soybean maturity gene haplotypes revealed by SNPViz analysis of 72 sequenced soybean genomes. PLoS One. 2014;9(4):e94150.
83. Prince SJ, Song L, Qiu D, Maldonado Dos Santos JV, Chai C, Joshi T, et al. Genetic variants in root architecture-related genes in a Glycine soja accession, a potential resource to improve cultivated soybean. BMC Genomics. 2015;16(1):132.
84. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(Database issue):D1178–86.
85. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
86. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001;29(1):37–40.
87. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A. 2004;101(30):11030–5.
88. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 2011;28(10):2731–9.
89. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 2001;17(12):1113–22.
90. Bülow L, Brill Y, Hehl R. AthaMap-assisted transcription factor target gene identification in *Arabidopsis thaliana*. Database. 2010;2010:034.
91. Joshi T, Fitzpatrick MR, Chen S, Liu Y, Zhang H, Endacott RZ, et al. Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. Nucleic Acids Res. 2014;42(Database issue):D1245–52.
92. Cannon EK, Cannon SB. Chromosome visualization tool: a whole genome viewer. Int J Plant Geno. 2011;2011:373875.
93. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34 suppl 2:W609–12.
94. Page RD. Visualizing phylogenetic trees using TreeView. In: Curr protoc bioinformatics. 2002. p. 6.2. 1–6.2. 15. vol. Chapter 6.
95. Liu C-M, Wong T, Wu E, Luo R, Yiu S-M, Li Y, et al. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics. 2012;28(6):878–9.
96. Van Dongen JT, Ammerlaan AM, Wouterlood M, Van Aelst AC, Borstlap AC. Structure of the developing pea seed coat and the post-phloem transport pathway of nutrients. Ann Bot. 2003;91(6):729–37.
97. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23(19):2633–5.
98. Milne I, Shaw P, Stephen G, Bayer M, Cardle L, Thomas WT, et al. Flapjack—graphical genotype visualization. Bioinformatics. 2010;26(24):3133–4.
99. Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305(3):567–80.