Genetic analysis of sucrose concentration in soybean seeds using a historical soybean genomic panel

Alexandra Ficht
University of Guelph
Robert W. Bruce
University of Guelph
Davoud Torkamaneh
University of Guelph
Christopher Grainger
University of Guelph
Milad Eskandari
University of Guelph
Istvan Rajcan (irajcan@uoguelph.ca)
University of Guelph https://orcid.org/0000-0001-5156-2482

Original Article

Keywords: GBS, SNP, QTL, GWAS, photosynthesis, plant growth, storage, nutrient assimilation

Posted Date: February 1st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-158915/v1

Version of Record: A version of this preprint was published at Theoretical and Applied Genetics on February 3rd, 2022. See the published version at https://doi.org/10.1007/s00122-022-04040-z.

Abstract

Soybean (*Glycine max* (L.) Merr) is a crop of global importance for both human and animal consumption, which was domesticated in China more than 6000 years ago. A concern about losing genetic diversity as a result of decades of breeding has been expressed by soybean researchers. In order to develop new cultivars, it is critical for breeders to understand the genetic variability present for traits of interest in their program germplasm. Sucrose concentration is becoming an increasingly important trait for the production of soy-food products. The objective of this study was to use a genome-wide association study (GWAS) to identify putative QTL for sucrose concentration in soybean seed. A GWAS panel consisting of 266 historic and current soybean accessions was genotyped with 76k genotype-by-sequencing (GBS) SNP data and phenotyped in four field locations in Ontario (Canada) from 2015 to 2017. Seven putative QTL were identified on chromosomes 1, 6, 8, 9, 10, 13 and 14. A key gene related to sucrose synthase (*Glyma.06g182700*) was found to be associated with the QTL found on chromosome 6. This information will facilitate efforts to increase the available genetic variability for sucrose concentration in soybean breeding programs and develop new and improved high-sucrose soybean cultivars suitable for the soy-food industry.

Key Message

Significant QTL for sucrose concentration have been identified using a historical soybean genomic panel, which could aid in the development of food-grade soybean cultivars.

Introduction

Soybean (*Glycine max* (L.) Merr) is a crop of global importance for both human and animal consumption and is used largely as a source of protein meal and vegetable oil in human diets (Hartman et al. 2011; Qiu et al. 2013). Historically, soybean has been a staple in Asian cuisine, but over time it has been integrated into both European and North American food markets. Soy-derived food products including tofu, soymilk, tempeh and natto have gained popularity over the past few decades due to an increase in popularity of plant-based diets (Lynch, Johnston and Wharton 2018). Soybeans provide a similar amount of protein as food from animal sources; however, they have lower amounts of saturated fat and little to no cholesterol, making them an ideal health food (Young 1991; Qiu et al. 2013). This has translated into an overall greater increase in the number of soy-food products in the global food sector.

Successful soybean food-grade cultivars are chosen based on physical traits, chemical traits and processing quality of the seed (Brar and Carter 1993, Rao et al. 2002). Sucrose concentration is a notable compositional trait that must be considered, especially because its target range differs depending on its intended end product. The range in seed sucrose concentration that is needed to meet the food-grade quality testing is between 6.0 to 8.0% (OSACC 2020). The *Ontario Soybean and Canola Committee* (OSACC) classifies cultivars based on seed sucrose concentration as low (< 6.4%), moderate (6.4 to 7.0%), and high (> 7.0%) (OSACC 2020).

Breeding for soybean seed sucrose concentration has received little attention, as the improvement of other agronomic and seed composition traits, including yield, seed oil, seed protein and disease resistance, are more important to commodity-grade soybeans. However, with shifts in demand, seed sucrose concentration is becoming an important chemical trait for breeders to focus on. Sucrose is the major product of photosynthesis and thus, a major sugar transported throughout higher plants. It has been shown to have many important functions including acting as a major substrate for sink metabolism, a main form of translocated carbon as well as a number of regulatory and integrative functions within the plant (Farrar et al. 2000). Sucrose metabolism has been shown to be linked to the metabolism of both organic and inorganic nitrogen and may also be an important factor for the balance of resource acquisition and allocation within and between plant organs (Farrar et al. 2000). Finally, the seed sucrose content of a plant is representative of a long-term balance between supply (e.g., photosynthesis) and demand (e.g., plant growth, storage and nutrient assimilation) (Farrar et al. 2000).

Sucrose metabolism is well defined in the literature and has, therefore, been a focus for plant biochemists. Sucrose synthase (SuSy) is a glycosyl transferase enzyme that is involved in sugar metabolism (Stein and Granot 2019). SuSy catalyzes the reversible cleavage of sucrose into fructose and either uridine diphosphate glucose (UDP-G) or adenosine diphosphate glucose (ADP-G; Stein and Granot 2019). Both cleaved products are then available for use by the plant, within many metabolic pathways including energy production, primary metabolite production and the synthesis of complex carbohydrates (Stein and Granot 2019). Previous studies have shown that the overexpression of *SUS* genes displayed increased growth, increased xylem area and cell-wall width and increased cellulose and starch content, indicating the potential for using *SUS* genes as candidate genes for the improvement of agronomic and chemical traits in crop plants (Stein and Granot 2019). Xu et al. (2019) identified 100 *SUS* genes in a number of higher plants, including soybean, and provided the phylogenetic relationship framework between the diverged *SUS* gene subfamilies (I, II and III). This information provides potential insight for its utility in marker-assisted selection.

Previous research on seed sucrose concentration has focused on elucidating quantitative-trait loci (QTL) using biparental populations (Maughan et al. 2000; Kim et al. 2005; Skoneczka et al. 2009; Zeng et al. 2014). However, there has been limited literature reporting on the detection of putative QTL for sucrose concentration using the genome-wide association study (GWAS) approach. Previous traditional QTL studies (i.e. linkage mapping) have identified significant QTL for sucrose concentration. Maughan et al. (2000) identified 17 sucrose-related QTL on chromosomes 5, 7, 8, 13, 15, 19 and 20. Kim et al. (2005) used RIL populations from a cross between ‘Keunolkong’ and ‘Shinpaldalkong’ and found four QTL for sucrose concentration on chromosomes 2, 11 and 19. Skoneczka et al. (2009) used F_2 derived populations from PI 87013 x PI 200508 and PI 243545 x PI 200508 and identified a QTL on chromosome 6. Finally, Zeng et al. (2014) crossed MFS-553 with PI 243545 and identified three novel QTL for sucrose concentration, which were located on chromosomes 5, 9 and 16.

The availability of high-throughput genotyping methods (e.g., genotyping-by-sequencing [GBS]) based on next-generation sequencing (NGS) technology allows for using GWAS as a method for the detection of

putative QTL related to soybean seed sucrose concentration. GWAS is a powerful tool for analyzing complex traits with a larger scope than traditional QTL identification strategies (Korte and Farlow 2013; Fang et al. 2017). Unlike biparental cross populations, the use of GWAS and a diversity panel captures allelic diversity and genetic background effects that may not have been uncovered otherwise (Heffner et al. 2009). The use of GWAS allows breeders to detect genomic regions associated with traits of interest and provides them with estimates regarding the size and direction of allelic effects (Abdel-Shafy et al. 2014; Contreras-Soto et al. 2017).

The objective of this study was to use GWAS to identify QTL related to sucrose concentration in soybean seed. The improved knowledge of particular QTL related to sucrose will allow for the integration of genetic variation into breeding germplasm leading towards the development of superior food-grade soybean cultivars with a focus on enhancing the global soy-food market.

Materials And Methods

Selection of Germplasm

Two populations of soybeans, representing breeding material, historical and elite cultivars indicative of decades of selection within the University of Guelph, Guelph Campus (GGC; 174 accessions) and University of Guelph, Ridgetown Campus (RC; 96 accessions) soybean breeding programs, were used to evaluate phenotypic diversity for seed sucrose concentration as described earlier (Bruce et al. 2019a; Bruce et al 2019b; and Bruce et al. 2020) (Table S1).

Field Trials

The GC population was grown at two field locations: (i) St. Pauls, Ontario from 2015 to 2017; and (ii) the Woodstock Research Station in Woodstock, Ontario from 2015 to 2017. The RC population was grown at two different locations: (i) Ridgetown, Ontario from 2015 to 2016; and (ii) Chatham, Ontario from 2015 to 2016. Trials were planted as nearest-neighbour randomized complete block designs (RCBD) with two replicates. Each four-row plot was planted with 500 seeds. Plots were 5 metres in length, with 40 cm row spacing. Each site was maintained using conventional management practices. Locations were harvested when every plot had reached full maturity.

Phenotypic Data Collection

Seed sucrose concentration was recorded for each plot. A Perten DA 7250 near-infrared reflectance (NIR) spectrometer (Perten Instruments, Hägersten, Sweden) was used to measure sucrose concentration using calibrations provided by Perten Instruments. A small tray of seeds from each plot were randomly sampled and hand-screened before each measurement was taken. Perten DA 7250 NIR estimates of sucrose concentration are imperfect, but unbiased (Table S2).

Statistical Analysis

Sucrose concentration data was examined in each environment (year location) to identify environments with poor performance. All environments in Guelph and Ridgetown had consistent seed trait performance; therefore, all data was kept for combined analysis. As a result of sampling errors in the Chatham 2016 environment, only a single-entry based seed sample was used and was therefore treated as a single block.

Plot data for sucrose concentration from St. Pauls and Woodstock, Ontario locations were spatially corrected using radial smoothing in PROC GLIMMIX in SAS 9.4 (SAS Institute, Cary, NC, USA) with the model including genotype and cov_sp as fixed effects and longitude (long) and latitude (lat; physical field positions) as random effects and an effect of cov_sp = spline (lat long). All further analyses used the spatially corrected data.

Both Chatham and Ridgetown locations were analyzed using a nearest-neighbour adjustment in Agrobase (Agronomix Software, 2019) to produce adjusted plot values. The model that was employed was , where is the plot value of the th cultivar of the th block for sucrose concentration, is the trial mean for sucrose concentration, is the effect of the th cultivar in the th block and is the residual error for the th cultivar in the th block, where the two adjacent plots are used in the calculation of the local trend.

Analyses of variance (ANOVA) of sucrose concentration within test locations were partitioned into fixed effects and random effects. Entry, Environment and Entry*Environment were used as the fixed effects and Block (Environment) was used as the random effect using the PROC GLIMMIX procedure in SAS version 9.4 (SAS Institute Inc., Cary, NC, USA) for a RCBD.

Combined analyses of variance were conducted over six environments (Chatham and Ridgetown from 2015 to 2016 and St. Pauls and Woodstock from 2015 to 2017) using the PROC GLIMMIX procedure in SAS version 9.4 (SAS Institute Inc., Cary, NC, USA). Least square means (LSMEANS) were calculated using PROC GLIMMIX procedure across single locations and combined location and year analyses. Pearson's correlation coefficients were calculated using the PROC CORR procedure performed between sucrose concentration LSMEANS and protein, yield and oil LSMEANS to determine any linear correlation.

Genotypic Data

DNA extraction was performed using the Qiagen DNeasy 96 Plant Kit (Toronto, Canada) following the manufacturer's protocol. The soybean panel was genotyped using the genotyping-by-sequencing (GBS) method, as described by Elshire et al. (2011) and Sonah et al. (2013) over three separate sequencing runs (Bruce et al. 2019a). Initially, 96 lines were sequenced in 2013 using the *Ape*KI enzyme on a single lane of HiSeq 2000 (Illumina, Inc.) at Genome Quebec, McGill University (Montreal, QC, Canada). This was followed by the sequencing of 168 genotypes in 2014, using *Msp*I/*Pst*I enzyme combination across three chips of Ion Torrent (Proton; Thermo Fisher Scientifc) at the Plateforme d'analyses génomiques [Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval (Quebec, QC, Canada), and finally, 96 genotypes were sequenced in 2016 with similar parameters to the 2014 sequencing, except that 35 of the genotypes were overlapping with the 2014 set of accessions to provide additional coverage and 29

overlapping with the 2013 set. Next-generation sequencing (NGS) sequence reads were processed through the Fast-GBS pipeline (Torkamaneh et al. 2017) and missing data imputation and integration was performed with BEAGLE v5 (Browning & Browning 2007) previously described by Torkamaneh and Belzile (2015) and Torkamaneh et al. (2018). In total, 76,549 genome-wide SNPs were used for this study. A minor allele frequency (MAF) filter of 0.05 and heterozygous SNP filter of 0.5 were applied prior to carrying out the GWAS. No missing data was present in the dataset. A principal component analysis (PCA) and phylogenetic tree analysis were performed in TASSEL v5 (Bradbury et al. 2007).

GWAS Analysis

GWAS analyses were performed with rMVP R package (Yin et al. 2020) using Fixed and random model Circulating Probability Unification (FarmCPU; Liu et al. 2016). The factored spectrally transformed linear mixed model (FaST-LMM) and efficient mixed model analysis (EMMA; Kang et al. 2008; Lippert et al. 2011) were used. A kinship matrix was calculated either using the VanRaden method (K) or the EMMA method (K*) to determine relatedness among individuals (Kang et al. 2008). The population structure was determined using fastSTRUCTURE (Raj et al., 2014) and a principal component analysis (PCA). Model taking into account kinship and PCA (P+K*) was found to provide the best fit. The negative log(1/p) was used to establish a significance threshold (Wang et al. 2012).

Results

Phenotypic analysis of seed sucrose concentration

The MC population breeding panel was evaluated for sucrose concentration in multi-environment trials during the 2015 to 2017 field seasons while the RC population breeding panel was grown during the 2015 and 2016 field seasons. The environment minimums, maximums and means for each environment are displayed in Table 1. The GC population displayed the highest average sucrose concentration at St. Pauls, Ontario in 2015 with a mean and standard error of 7.37 ± 0.067. The lowest average sucrose concentration was observed in Woodstock, Ontario in 2017 with an overall mean seed sucrose concentration of 6.16 ± 0.067. The RC population displayed the highest average sucrose concentrations were observed at both Ridgetown and Chatham, Ontario locations in 2015 with a mean seed sucrose concentration of 6.52 ± 0.115 and 6.52 ± 0.068, respectively. The lowest average sucrose concentration in a single environment was observed at Chatham, Ontario in 2016 with a mean seed sucrose concentration of 6.13 ± 0.062.

The combined GC and RC populations showed a normal distribution (Shapiro-Wilk, $p<0.05$) for seed sucrose concentration (Figure 1A), with an overall mean of 6.83 ± 0.9. A combined analysis of variance was carried out for the combined GC and RC populations displaying a significant difference among genotypes, environments and genotype*environment effects in the populations (Tables S3 and S4). Broad sense heritability (H^2) was calculated for sucrose concentration and found a moderate to high H^2, 0.76 and 0.78 for GC and RC panel, respectively (Table 2).

Relationship between traits

The relationship between sucrose concentration and other value-added and agronomic traits (seed protein, oil and yield) was also studied in both the GC and RC populations (Tables 3 and 4). The Woodstock yield data for 2017 was dropped due to error. Analysis by year revealed significant negative relationships between oil and proteins in across both years in both populations (Tables 3 and 4). A significant negative relationship between sucrose and protein was observed in across all years at the $p < 0.001$ level (Table 3). A positive relationship was observed between sucrose concentration and oil; however, this relationship was not significant. A significant positive relationship between sucrose concentration and yield was observed for the GC population in 2015, 2016 and 2017 (Table 3).

Genotypic analysis of seed sucrose concentration

In total, 76,549 high-quality genome-wide GBS-SNPs were identified in this population consisting of 266 soybean genotypes. A PCA and phylogenetic tree clustering were carried out using 76k SNPs (Figures 1B & C) indicated genetic admixture within the panel, without any tight or distinct subsets between the GC and RC populations, due to shared founder germplasm. A single best linear unbiased estimator (BLUE) for sucrose concentration and 76k GBS-SNPs data were used as input for the GWAS. The GWAS analysis was performed using FarmCPU model, and population structure (P) and cryptic relatedness (K*) were incorporated as covariates to reduce false positive signals. Using this approach, we identified seven QTL ($-\log_{10} P \geq 4.5$) for seed sucrose concentration on chromosomes 1, 6, 8, 9, 10, 13 and 14 (Figure 2A; Table 5). Of these, qSUC.1.39, qSUC8.22 or qSUC14.26 were novel QTL, which have not been reported previously in those genomic regions on chromosomes 1, 8 and 14, respectively. One of the most intriguing features of this study is that almost all alleles related to these QTL were frequently (~28%) present in this population. The estimated allele effects of the putative QTL indicate that they control close to 30% of the phenotypic variation within this population (Table 5).

Candidate genes

We used the soybean public database (SoyBase (SoyBase 2020)) and soybean reference genome (Williams 82) annotation (Wm82.a2.v1) to identify candidate genes for sucrose accumulation. To search for candidate genes, the QTL flanking regions were set up as 100kb on either side of the QTL peak. There were no previously identified QTL or genes in the region associated with qSUC.1.39, qSUC8.22 or qSUC14.26.

The gene associated with qSUC.6.15 is *Glyma.06g182700,* is located approximately 361bp upstream of the SNP peak (15682704). It is annotated as a protein encoding carbonic anhydrase. This gene has been shown to have an indirect role in the sucrose metabolic pathway. Gene expression data provided by Severin et al. (2010) showed that *Glyma.06g182700* is highly expressed in soybean root nodules.

The gene associated with qSUC.9.2 is *Glyma.09g035700*, with the peak (2989052) falling within the gene itself. *Glyma.09g035700* is annotated as iron-sulfur binding protein. This gene plays a role in assisting

with the oxidation-reduction reactions of electron transport in both mitochondria and chloroplasts (Balk and Pilon 2011). Gene expression data provided by Severin et al. (2010) showed that *Glyma.09g035700* is highly expressed in the soybean young leaf and flower, with moderate expression in soybean pods, 7 and 10 to 13 days after fertilization (DAF) and soybean seeds from 14 to 35 DAF.

The gene associated with qSUC.10.5 is *Glyma.10g060200*, with the peak (5593982) falling within the gene itself. It is annotated as glutamate-ammonia ligase. This gene is involved with the production of glutamine in the glutamine biosynthetic pathway (Miflin Wallsgrove and Lea 1981). Gene expression data provided by Severin et al. (2010) showed that *Glyma.10g060200* is highly expressed in the soybean flower and moderate expression in the young leaf and pod shell.

The gene associated with qSUC.13.17 includes *Glyma.13g070500*, with the peak (17093380) falling within the gene itself. It is annotated as a protein with N-terminal bromo-adjacent homology (BAH) and transcription elongation factor S-II (TFS2N) domains (SoyBase 2020). It is shown to be involved in the miRNA pathway and involved in translational repression (TAIR 2020). Gene expression data provided by Severin et al. (2010) showed that Glyma.13g070500 is highly expressed throughout the entire plant during the plant's development.

Discussion

The majority of soybean breeding programs have focused on the improvement of commodity-related traits, including yield, seed protein and seed oil concentration. Increased demand for soy-food products has shifted the focus of some soybean breeding programs to include value-added, food-grade traits, such as seed sucrose concentration. The identification of sucrose-related QTL would aid the development of high-sucrose soybean cultivars for the food-grade market through marker-assisted selection.

Fewer than ten other studies have sought to identify sucrose-related QTL in soybean, with each utilizing bi-parental RILs populations (Maughan et al. 2000; Kim et al. 2005; Skoneczka et al. 2009; Zeng et al. 2014). While these studies have identified over 25 sucrose-related QTL, these loci have limited applicability in marker-assisted selection schemes using different genetic backgrounds. The improvement of seed sucrose concentration became a priority for the University of Guelph soybean breeding programs, and it became necessary to identify novel sucrose-related QTL in a vast array of genetic backgrounds associated with these programs. Before the current study, GWAS had not been used to analyze genetic data for putative sucrose-related QTL.

Overall, seven putative QTL for seed sucrose concentration were identified. Of the seven QTL, both qSUC.13.17 (chromosome 13) and qSUC.9.2 (chromosome 9) were validated by previously identified QTL (Panthee et al. 2006; Zeng et al. 2014). A previously identified QTL for seed sucrose concentration, Seed sucrose 4-2, was shown to explain approximately 10% of the variation in seed sucrose concentration (Zeng et al. 2014). It can be assumed that this overlapping QTL contributed to 10% of the phenotypic variation. qSUC.13.17 was shown to overlap with two previously identified QTL, Seed Cys 2-3 and Seed Met 2-1 (Panthee et al. 2006). The remaining five QTL that were identified are considered novel.

Candidate genes for each QTL were identified. A number of the genes that play roles in general biosynthetic pathways including the glutamine pathway. However, one candidate gene, *Glyma.06g182700* encode carbonic anhydrases, has been shown to play a role in the sucrose metabolic pathway specifically (Kavroulakis et al. 1999). A similar gene and function has previously been reported in pigeonpea (*Cajanus cajan*; Dutta et al. 2011).

Although putative QTL for seed sucrose concentration were found, they were limited in number and effect size. The broad sense heritability (H^2) was calculated for both populations and was found to be moderate to high, with the GC population having a H^2 of 0.76 and the RC population having a H^2 of 0.78. The moderate to high heritability for seed sucrose indicated that approximately ~76% of the phenotypic variation for seed sucrose concentration was controlled genetically, inferring that phenotypic selection has the ability to increase genetic gain in the breeding program. A potential reason for locating no more than seven QTL may be the result of the method of determining seed sucrose levels. NIR estimation of sucrose concentration though unbiased, could be variable (Table S1). This indicates that we had somewhat limited power to detect QTL or map them to precise locations in the genome. Moreover, the limited range in seed sucrose concentration is the direct result of selecting for other important agronomic and quality traits. As most of the breeding focus was geared towards increased yield and seed protein or oil concentration, the available genetic diversity was lost. In order to regain variation in seed sucrose concentration, the introduction or introgression of high-sucrose parental genotypes should be explored.

Currently, there is minimal research available for QTL for seed sucrose concentration; however, with the increased interest in soy-food products, sucrose is becoming a more important trait for breeders. Sucrose is an important quality trait for soy-food products, with an acceptable range of approximately 6.0 to 8.0% seed sucrose concentration. More research is needed to provide soybean breeders with confirmed QTL for seed sucrose concentration to allow for their use in marker assisted selection. After decades of selection for agronomic traits like yield and seed protein concentration, the available variation in seed sucrose in soybean breeding germplasm may have been reduced compared to ancestral cultivars. The findings of these putative QTL provide a foundation for the elucidation of useful QTL for the future development of food-grade soybean cultivars.

Declarations

Author Contributions

IR conceptualized the study, obtained funding for the study and contributed to the design, analysis and interpretation and writing of the manuscript. RB collected data in the 2015 and 2016 seasons and contributed to the analysis and interpretation of the manuscript. DT contributed to data analysis and interpretation of the manuscript. CMG contributed to the methodology used, DNA sampling, writing of the NSERC-CRD grant that funded the research and editing of the manuscript. ME contributed to the conceptualization of the research, genomic panel assembly, data collection and manuscript editing. AF

collected the data in the 2016 and 2017 seasons, carried out the experiments, analyzed the data and wrote the manuscript as a part of her M.Sc. thesis at the University of Guelph, Guelph, ON, Canada.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We would like to acknowledge funding support from the Natural Sciences and Engineering Research Council of Canada Collaborative Research and Development Grant (NSERC CRD) program and our industry partners including Grain Farmers of Ontario, Secan and Huron Commodities Inc. All of the mentioned funding is part of the NSERC CRD grant no. CRDPJ 447948 – 13. We would also like to acknowledge the staff in Dr. Rajcan's research program in Guelph, ON, including Colbey Templeman, Yesenia Salazar, Martha Jimenez, Mei Wang and Lin Liao, as well as Bryan Stirling in Dr. Eskandari's lab at Ridgetown, ON.

References

Abdel-Shafy H, Bortfeldt RH, Tetens J, Brockmann G (2014) Single nucleotide polymorphism and haplotype effects associated with somatic cell score in German Holstein cattle. Genet Sel Evol 46:35

Agronomix Software, Inc (1999) Agrobase 99 user's guide and reference manual. Agronomix Software, Inc., Winnipeg, MB. 428 pp

Balk J, Pilon M (2011) Ancient and essential: the assembly of iron-sulfur clusters in plants. Trend Plant Sci 16:218-226

Brar GS, Carter TE (1993) Soybean *Glycine max* (L.) Merrill. *In* G. Kalloo and B.O. Bergh (ed.) Genetic improvement of vegetable crops. Pergamon Press, New York. Pp. 427-463

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097

Bruce RW, Torkamaneh D, Grainger C, Belzile F, Eskandari M, Rajcan I (2019a) Genome-wide genetic diversity is maintained through decades of soybean breeding in Canada. Theor Appl Genet 132:3089-3100.

Bruce, R, Torkamaneh, D, Grainger, C, Belzile, F, Eskandari, M, Rajcan I. 2019. Genome-wide genetic diversity is maintained through decades of soybean breeding. Theor. Appl. Gen. 132:3089–3100

Bruce, RW, Torkamaneh, D, Grainger, CM, Belzile, F, Eskandari, M and Rajcan I. 2020. Haplotype diversity underlying quantitative traits in Canadian soybean breeding Germplasm. Theor. Appl. Genetics 133:1967–1976

Bruce RW, Torkamaneh D, Grainger C, Belzile F, Eskandari M, Rajcan I (2020) Haplotype diversity underlying quantitative traits in Canadian soybean breeding germplasm. Theor Appl Genet 133:1967-1976

Contreras-Soto RI, Mora F, Rott de Oliveira MA, Higashi W, Scapim CA, Schuster I (2017) A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. PLoS ONE 12:e0171105

Dutta S, Kumawat K, Singh BP, Gupta DK, Singh S, Dogra V, Gaikwad K, Sharma TR, Raje RS, Bandhopadhya TK, Datta S, Singh MN, Bashasab F, Kulwal P, Wanjari KB, Varshney RK, Cook DR, Singh NK (2011) Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea *Cajanus cajan* L. Millspaugh. BMC Plant Biol 11:17

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379

Fang C, Ma Y, Wu S, Zhi L, Zheng W, Yang R, Hu G, Zhou Z, Yu H, Zhang M, Pan Y, Zhou G, Ren G, Du W, Yan H, Wang Y, Han D, Shen Y, Liu S, Liu T, Zhang J, Qin H, Yuan J, Yuan X, Kong F, Liu B, Li J, Zhang Z, Wang G, Zhu B, Tian Z (2017) Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol 18:161

Farrar J, Pollock C, Gallagher J (2000) Sucrose and the integration of metabolism in vascular plants. Plant Sci 154:1-11

Hartman GL, West ED, Herman TK (2011) Crops that feed the world 2. Soybean worldwide production, use, and constraints caused by pathogens and pests. Food Security 3:5-17

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1-12

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723

Kavroulakis N, Flemetakis E, Aivalakis G, Katinakis P (1999) Carbon metabolism in developing soybean root nodules: The role of carbonic anhydrase. Mol Plant Microbe Interact 13:14-22

Kim HK, Kang ST, Cho JH, Choung MG, Suhd Y (2005) Quantitative trait loci associated with oligosaccharide and sucrose contents in soybean (*Glycine max* L.). J Plant Biol 48:106-112

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. Nat Methods 8:833-835

Liu XL, Huang M, Fan B, Buckler ES, Zhang ZW (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767

Lynch H, Johnston C, Wharton C (2018) Plant-based diets: Considerations for environmental impact, protein quality and exercise performance. Nutrients 10:1841

Maughan PJ, Maroof MAS, Buss GR (2000) Identification of quantitative trait loci controlling sucrose content in soybean (Glycine max). Mol Breed 6:105-111

Miflin BJ, Wallsgrove RM, Lea PJ (1981) Glutmaine metabolism in higher plants. Curr Top Cell Reg 20:1-43

OSACC (2020) Ontario soybean and canola committee. Available online: http://www.gosoy.ca. Accessed 20 October 2020

Panthee D, Pantalone V, Sams C, Saxton A, West D, Orf J, Killam A (2006) Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. Theor Appl Genet 112:546:553

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909

Qiu LJ, Zing LL, Guo Y, Wang J, Jackson SA, Chang RZ (2013) A platform for soybean molecular breeding: the utilization of core collections for food security. Plant Mol Bio 83:41-50

Rao MSS, Mullinix BG, Rangappa M, Cebert E, Bhagsari AS, Sapra VT, Joshi JM, Dadson RB (2002) Genotype x environment interactions and yield stability of food-grade soybean genotypes Agron J 94:72-80

SAS Institute 2013. SAS® 9.4. SAS Institute. Cary, NC

Severin AJ, Woody JL, Bolon YT, Jospeh B, Diers BW, Farmer AD, Muehlbauer GJ, Nelson RT, Grant D, Specht JE, Graham MA, Cannon SB, May GD, Vance CP, Shoemaker RC (2010) RNA

Seq atlas of *Glycine max:* A guide to the soybean transcriptome. BMC Plant Biol 10:160

Skoneczka JA, Saghai Maroof MA, Shang C, Buss GR (2009) Identification of candidate gene mutation associated with low stachyose phenotype in soybean line PI200508. Crop Sci 49:24-255

Sonah H, Bastien M, Iquira E, Tardivel A, Legare G, Boyle B, Normandeau E, Laroche J, Larose S, Jean M, Belzile F (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8:e54603

Stein O, Granot D (2019) An overview of sucrose synthases in plants. Front Plant Sci 10:1-14

Torkamaneh D, Belzile F (2015) Scanning and filling: ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. PLoS ONE 10:e0131533

Torkamaneh D, Laroche J, Tardivel A, O'Donoughue L, Cober E, Rajcan I, Belzile F (2017)

Comprehensive description of genomewide nucleotide and structural variation in short season soya bean. Plant Biotechnol J 16:749:759

Torkamaneh D, Boyle B, Belzile F (2018) Efficient genome-wide genotyping strategies and data integration in crop plants. *Theor. Appl. Genet.,* 131:499-511

Wang M, Yan J, Zhao J, Song W, Zhang X, Xiao Y, Zheng Y (2012) Genome-wide association study (GWAS) of resistance to head smut in maize. Plant Sci 196:125-131

Xu X, Yang Y, Liu C, Sun Y, Zhang T, Hou M, Huang S, Yuan H (2019) The evolutionary history of the sucrose synthase gene family in higher plants. BMC Plant Biol 19:566

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Xiaolei L (2020) rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome

wide association study. bioRxiv doi:10.1101/2020.08.20.258491

Young VR (1991) Soy protein in relation to human protein and amino acid nutrition. J Am Diet Assoc 91:828-835

Yu J, Pressoir G, Briggs WH, Vroh BI, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208

Tables

Table 1. Variation for sucrose concentration in the University of Guelph, Guelph Campus (GC) and University of Guelph, Ridgetown Campus (RC) soybean breeding program populations. The MC population was grown in two locations: (1) St. Pauls and (2) Woodstock from 2015 to 2017. The RC population was grown in two different locations: (1) Chatham and (2) Ridgetown in 2015 and 2016.

GC Population	Environment	Minimum	Maximum	Mean (se)[a]	CV (%)
	St. Pauls 2015	5.25	8.93	7.37 (0.067)	10.7
	St. Pauls 2016	1.48	8.84	6.92 (0.063)	11.7
	St. Pauls 2017	5.29	9.31	7.19 (0.067)	11.9
	Woodstock 2015	4.96	8.67	6.76 (0.061)	10.6
	Woodstock 2016	2.06	9.02	7.18 (0.069)	12.4
	Woodstock 2017	3.90	8.72	6.16 (0.067)	14.0
	Combined	1.48	9.31	6.75 (0.061)	14.7
RC Population	**Environment**	**Minimum**	**Maximum**	**Mean (se)[a]**	**CV (%)**
	Chatham 2015	4.81	8.77	6.52 (0.068)	10.1
	Chatham 2016	4.41	7.34	6.13 (0.062)	9.81
	Ridgetown 2015	2.24	9.01	6.52 (0.115)	16.3
	Ridgetown 2016	1.60	8.67	6.43 (0.118)	16.9
	Combined	1.60	9.01	6.44 (0.093)	14.1

[a] Data represent the standard error (a = 0.05).

Table 2. Broad-sense heritability of seed sucrose concentration in two soybean genomic panels: MCPOPn evaluated in six environments (STPL15, STPL16, STPL17, WDSK15, WDSK16, WDSK17) and RCPOPn evaluated in four environments (CHAT15, CHAT16, RT15, RT16).

	Sucrose
GCPOPn	0.7635
RCPOPn	0.7806

Table 3. Pearson's correlations between sucrose concentration and agronomic traits in a diversity panel developed at the University of Guelph, Gueph Campus (GC POPn). St. Pauls and Woodstock locations were pooled across the 2015 to 2017 growing seasons. Data represent the Pearson correlation coefficients.

Environment	Trait	Oil	Protein	Yield
2015		0.07	-0.66***	0.41***
2016	Sucrose	-0.04	-0.55***	0.25***
2017		-0.26***	-0.31***	0.15*
2015			-0.66***	0.03
2016	Oil		-0.67***	0.14**
2017			-0.51***	0.06
2015				-0.36***
2016	Protein			-0.20**
2017				-0.05

* Significant at p < 0.05. ** Significant at p < 0.01. *** Significant at p < 0.001.

Table 4. Pearson's correlations between sucrose concentration and agronomic traits in a diversity panel developed at the University of Guelph Ridgetown campus (RC POPn). Chatham and Ridgetown locations were pooled across the 2015 and 2016 growing seasons. Data represent the Pearson correlation coefficients.

Environment	Trait	Oil	Protein	Yield
2015	Sucrose	0.11*	-0.10	-0.11*
2016		-0.03	-0.05	-0.21**
2015	Oil		-0.36***	0.09
2016			-0.37***	-0.02
	Protein			0.19***
2016				-0.12*

* Significant at p < 0.05. ** Significant at p <0.01. *** Significant at p < 0.001.

Table 5. Putative sucrose QTL identified through analysis in FarmCPU of a pooled 266 soybean genotype diversity panel developed at the University of Guelph Campus (GC) and University of Guelph Ridgetown Campus (RC). The lines were grown in four locations, Chatham and Ridgetown from 2015 to 2016 and St. Pauls and Woodstock, Ontario from 2015 to 2017.

SNPiD	Chromosome	Position	P-value	Alleles	Allele Effect[a]
qSUC.1.39	1	39218553	8.27E-10	70% A : 30% G	A = 7.15; G = 6.46
qSUC.6.15	6	15682704	1.74E-06	64% G : 36% A	A = 6.77; G = 6.84
qSUC.8.22	8	22891546	4.42E-08	74% G : 26% A	A = 6.80; G = 6.83
qSUC.9.2	9	2989052	4.46E-06	78% T : 22% C	T = 6.89; C = 6.63
qSUC.10.5	10	5593982	3.24E-05	58% A : 42% G	A = 6.66; G = 7.08
qSUC.13.17	13	17093380	6.53E-06	70% G : 30% A	A = 6.84; G = 6.90
qSUC.14.26	14	26432629	3.58E-07	91% A : 9% G	A = 6.89; G = 6.51

[a] Mean seed sucrose concentration per allele.

Figures

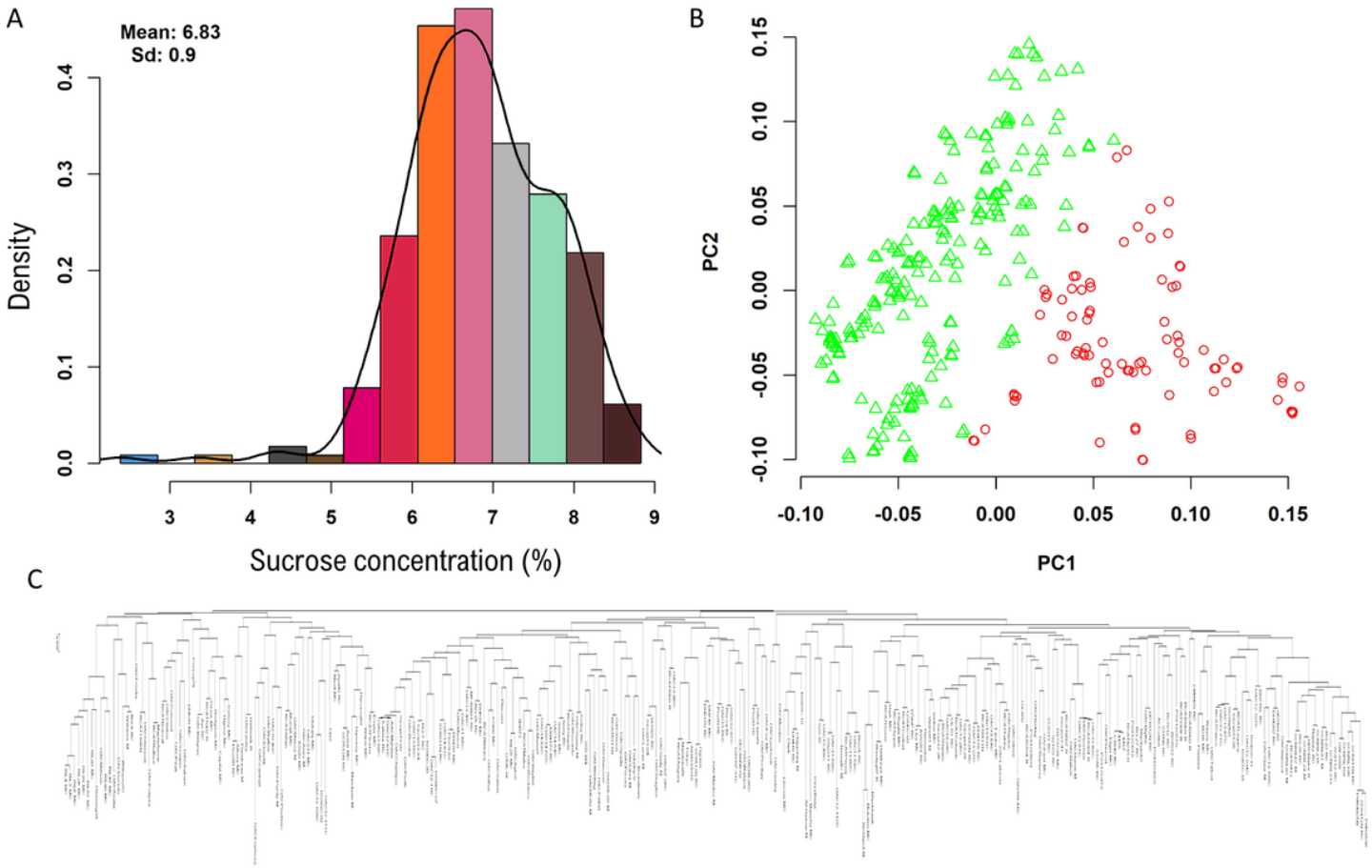

Figure 1

Phenotypic analysis of seed sucrose concentration in the University of Guelph, Guelph Campus (GC) population from 2015 to 2017 and Ridgetown Campus (RC) population from 2015 to 2016. (A) Distribution of seed sucrose concentration variation in the combined populations. (B) Principal component analysis of seed sucrose concentration in the combined GC and RC populations. (C) Maximum likelihood phylogenetic tree of seed sucrose concentration showing overlap between the GC and RC populations. The green lines represent OAC lines from the GC population and the red lines represent RCAT lines from the RC population.

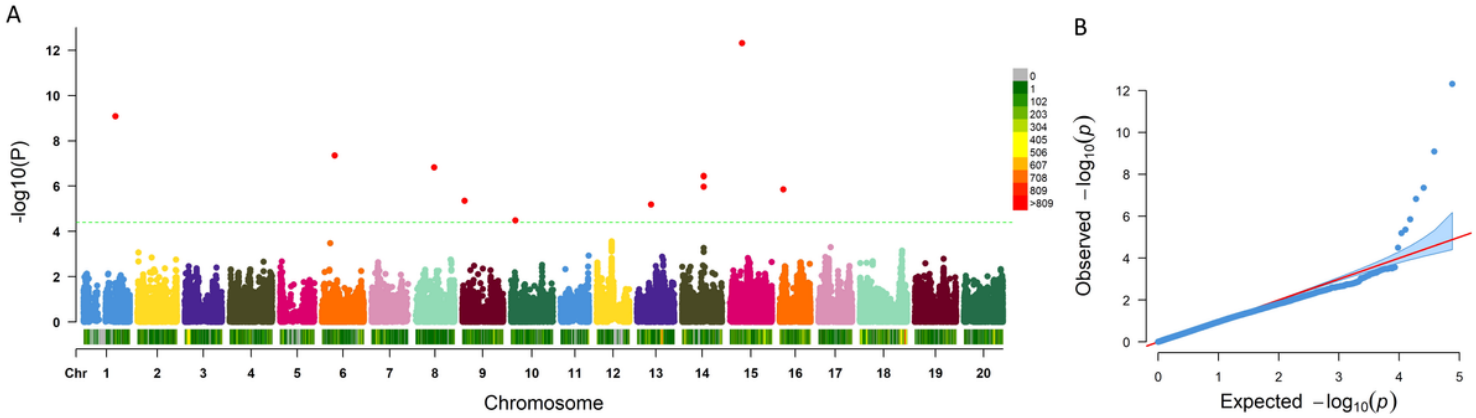

Figure 2

Genome wide association study for seed sucrose concentration in the University of Guelph, Guelph Campus (GC) and Ridgetown Campus (RC) populations. (A) A Manhattan plot showing 7 putative QTL for seed sucrose concentration. (B) The Normal Q-Q plot for seed sucrose concentration in the GC and RC populations.

Supplementary Files

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

- SucroseQTLFileS1.xlsx
- SucroseQTLFileS2.xlsx
- SupplementaryMaterial.docx