Genome-wide SNP detection in Darjeeling tea: unravelling genetic structure, functional impact and trait associations

Anjan Hazra¹, Rakesh Kumar², Chandan Sengupta³, Sauren Das^{1*}

¹Agricultural and Ecological Research Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India

²Darjeeling Tea Research and Development center, Kurseong, West Bengal 734203, India

³Department of Botany, University of Kalyani, Nadia 742325, India

*Corresponding author: sauren@isical.ac.in

Abstract

Genotyping by sequencing and identification of functionally relevant nucleotide variations in crop accessions are the key steps to unravel genetic control of desirable traits. In the scope of present work, elite cultivars of Darjeeling tea were undergone SNP genotyping by double-digest restriction site associated DNA sequencing method. This study reports a set of 54,206 high-quality SNP markers discovered from ~10.4 GB sequence data, encompassing 15 chromosomes of reference tea genome. Genetic relatedness among the accessions conforms to the analyses of Bayesian clustering, UPGMA, and PCoA methods. High percent of heterozygous loci in the majority of the cultivars reflect their ‘hybrid’ ancestry as reported earlier. Genomic positions of the discovered SNPs and their putative effect on annotated genes designated a thoughtful understanding of their functional aspects in tea system biology. A group of 95 genes was identified to be affected by high impact variants, most of them are involved in signal transduction, biosynthesis of secondary metabolite, transcriptional and translational regulation. Genome-wide association analyses of 21 agronomic and biochemical phenotypes resulted in trait-linked polymorphic loci with strong confidence ($p < 0.05$ and 0.001). The selection of significant marker-trait associations with the Bonferroni correction threshold retained a set of 57 SNPs distributed across 14 chromosomes to be linked with eight phenotypic traits. High impact and trait-associated nucleotide polymorphisms perceived in this study can be exploited in worldwide germplasm of contrasting origin to depict their heritability and to unlock their potentiality in marker-assisted breeding.

Keywords

Double-digest restriction site associated DNA; Darjeeling tea; Genotyping-by-Sequencing; Genome-wide association studies; Single nucleotide polymorphisms; SNP effects

Introduction

Tea, the globally second most-consumed drink after water, is popular for its irreplaceable refreshing value and potential health benefits. Following industrial manufacturing of the fresh leaves from perennial evergreen *Camellia sinensis* plants, processed tea possesses a certain distinctive set of secondary metabolites that contribute to its pleasant aroma, taste, and other wellbeing attributes (Hazra et al., 2019; Xia et al., 2020). Although the plant itself grows in diverse habitats, the coziness and quality vary with the geographical location and cultivar origin (Firmani et al., 2019; Hazra et al., 2018a). Since its inception, Darjeeling tea, in particular, is well known for its premium quality and brand value in the world beverage market. The temperature-moisture combination of Darjeeling hills prevailing throughout the year, the genetic background of the planted bushes, and overall the undergoing processing stages have been ascribed to the generation of a majestic flavor associated with the superior quality of this tea (Das, 2006). Owing to the excellence and physiognomies, Darjeeling tea as the first Indian product, achieved Geographical Indications tag in 2004–2005 (Das, 2006) so that the cultivation and production of the same are maintained and restricted to the geographical territory of Darjeeling hills. The report states that the production of Darjeeling tea contributes only 1% of the total Indian tea production, however, in terms of economic paybacks, it supersedes all others since last considerable period (Hazra et al., 2018b). Nonetheless, the recent gradual changes in temperature and rainfall patterns in Darjeeling hills imposing an alarming threat to the legendary Darjeeling tea industry leading to deterioration in overall production (Singh, 2019).

Being a perennial crop and evolved to thrive in diverse ecological conditions, specific cultivars of tea plants with an anticipated genetic makeup, possess unique traits of climate-smart attributes. In addition to various stress resilience, they can also serve as an efficient and sustainable natural exhauster by capturing and storing CO₂ in their dense green coverage along mountain slopes and valleys (Phukan et al., 2018). Hence, the selection and utilization of vigor cultivars with desirable phenotypes are crucial toward the prospective crop improvement program in tea. Development of trait-associated molecular markers has been adopted for tea in the recent past (Hazra et al., 2018a; Mukhopadhyay et al., 2016). Due to several bottlenecks in the life cycle, conventional association mapping has been found difficult in tea (Hazra et al., 2020a). Hence, simultaneous genotyping and phenotyping of

contrasting core accessions and subsequent marker-trait association studies were suggested as a convenient choice to overcome the challenges (Hazra et al., 2019; Hazra et al., 2018a).

Single nucleotide polymorphisms (SNPs) are considered as a superior option among the molecular markers because of their high frequency in genomes and expedient detectability using high-throughput approaches (Guajardo et al., 2020; Mammadov et al., 2012). SNPs can be accepted as key factors for the generation of phenotypic variation and their effect on functional alteration of genes enlightens the functional genomics of an organism (Hirakawa et al., 2013). Genotyping-by-Sequencing (GBS) applies next-generation sequencing approaches to enable both the *de-novo* and reference-based genotyping of numerous SNPs in a diverse set of samples (Elshire et al., 2011). Among the several types of high-throughput SNP genotyping methods discovered in recent years, double-digest restriction site associated DNA (ddRAD) sequencing is one of the most convenient, cost-effective, and widely used platforms (Peterson et al., 2012; Ray and Satya, 2014). This genome-wide approach consecutively involves (i) low- and high-frequency cuts to digest the DNA, (ii) ligation of a barcoded adapter to one restriction site and a common adapter to the other, (iii) pooling, (iv) size selection, (v) library enrichment, and (vi) introduction of a second barcode in the form of an Illumina index to increase multiplexing (Scheben et al., 2017). The ddRAD generated sequence data have been employed for various plant breeding approaches like linkage mapping (Davik et al., 2015), QTL analyses (Chen et al., 2017; Laila et al., 2019; Sasai et al., 2019), and genome-wide association studies (Jaiswal et al., 2019; Luo et al., 2019; Raggi et al., 2019; Sudan et al., 2019). Assessment of the genetic diversity of different plant species, such as apple (Ma et al., 2017), orchid (Roy et al., 2017), *Prunus* (Guajardo et al., 2020), Sesame (Basak et al., 2019) and onion (Lee et al., 2018) have also been carried out through ddRAD technology. In tea, ddRAD based SNP genotyping was applied for the determination of genetic diversity and population structure among the germplasm accessions and their wild relatives from China and Japan (Yamashita et al., 2019; Yang et al., 2016).

Darjeeling tea cultivars of diverse genetic makeup that carrying contrasting phenotypic aptitudes are being maintained at the experimental field of Darjeeling Tea Research and Development Centre for research and trial purposes. Dissecting the genetic background of the desirable phenotypes underlying these accessions is needed to excavate their breeding potential. Extensive phenotyping and simultaneous genotyping are considered the most suitable option for attending the resolution of efficient breeding (Raggi et al., 2019).

Sequence-based genetic characterizations, especially with SNP markers are too inadequate to uncover the potential of these genomic resources in India. To the best of our knowledge, ddRAD based, or any other GBS for Darjeeling tea cultivars have not been reported so far. In this study, the core collection of 23 elite accessions for the cultivation of Darjeeling tea were taken into account for the ddRAD-seq assisted genotyping approach. Identified SNPs distributed throughout the tea genome were then assessed to derive the levels of molecular diversity and genetic structure among the accessions, extrapolation of SNP effects, and identification of SNP-carrying genes. Genome-wide association studies with the high confidence variant sites were accompanied for the major agronomic traits of tea, to harvest desirable trait-associated loci. SNPs with potential functional impact and/or phenotypic trait relations were presented for possible implementation in the marker-assisted selection and thereby forthcoming crop improvement programs of tea.

Materials and Methods

Plant materials

A total of twenty-three elite tea cultivars released for plantation at the various slopes of Darjeeling hills were chosen for this study. All these accessions are being maintained at Darjeeling Tea Research and Development Center (DTRDC), Kurseong, with uniform agro-ecological conditions for research and breeding purposes (Hazra et al., 2020a; Hazra et al., 2020b). Most of the cultivars are commercial hybrid, generated through preferential breeding techniques among the wild and domesticated types of tea lineages. Uniqueness in yield and flavor potential of the commercial accessions were described earlier, through evaluation in long-term field trial (Singh, 1992). These include the widely cultivated varieties by Darjeeling tea planters for their excellent aroma traits (AV-2, K-1/1, and B-668), very high yield (HV-39 and T-78), or drought-tolerant characteristics (RR/17/144). Fresh leaf and bud samples were harvested from each of the studied cultivars, kept in frozen condition, transferred to the laboratory, and stored in -80°C refrigerator until further use.

Genotyping by sequencing and computational processing

Genomic DNA extraction from harvested young leaves of each of the cultivars was performed using the DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's instructions. The quality and quantity of the isolated DNA were examined by 1% agarose gels (Figure S1) and a Qubit Fluorometer v2.0 (Thermo Scientific, USA) respectively. For

GBS, a double digest restriction associated DNA (ddRAD) sequencing protocol was employed. A total of 1 µg of intact genomic DNA for each sample was digested using *SphI* and *MluI* restriction enzymes. Digested products were cleaned up using Ampure XP beads (Beckman Coulter Genomics) followed by ligation of P1 (barcoded) and P2 adaptors using T4 DNA ligase. Following subsequent pooling and cleanup, the size selection of the ligated products was carried out by 2% agarose gel electrophoresis. PCR amplification was done to enrich and add the Illumina specific adapters and flow cell annealing sequences. QC of the prepared ddRAD libraries for each of the samples was checked using a Bio Analyser (Agilent Technologies, USA). All the 23 ddRAD libraries with satisfactory QC profiles were then sequenced on an Illumina HiSeqX 10 platform (AgriGenome Labs Pvt. Ltd., India) using a 2×150 bp paired-end read module. Raw sequence reads were demultiplexed and filtered based on RAD tags at both ends of the pair. The quality of paired-end raw reads was examined in FastQC (Andrews, 2010). Filtered reads were pre-processed in Trimmomatic-0.33 (Bolger et al., 2014) for removal of the Illumina adapters and 5' and 3' base trimming. Subsequently, the remaining clean reads were subjected to paired-end alignment against the tea plant reference genome (Xia et al., 2020) using Bowtie v2.3.4.3 (Langmead et al., 2009) with default parameters. The ddRAD sequencing generated raw reads has been deposited in the NCBI Sequence-Read Archive (SRA) database.

Genome-wide SNP detection

Reference genome aligned sample-specific BAM files were sorted by coordinate, de-duplicated, and assigned appropriate read groups with the help of SAMtools (Li et al., 2009) and Picard Toolkit (<http://broadinstitute.github.io/picard/>). Based on the preprocessed BAM files, the SNP calling was performed using FreeBayes (Galaxy Version 1.3.1) (Garrison and Marth, 2012) on a galaxy interface (Afgan et al., 2018). The bi-allelic SNPs were filtered at minimum read depth 10 and de-complexed using the VCFlib toolkit (Garrison and Marth, 2012). Identified polymorphic SNPs with 10X read coverage were further subjected to filtering for minor allele frequency (MAF) ≥ 0.05 .

Estimation of SNP stats, heterozygosity and population structure

SNP densities across 15 chromosomes of the tea plant were represented by histogram and heatmap using Circos program v0.69-8 (Krzywinski et al., 2009) and plotted in 1 Mbp window size. Minor allele frequency and Tajima's D distribution of the polymorphic sites, frequencies of SNP across chromosomes, and transition-transversion ratio were estimated via

the SNIplay v3.0 tool (Dereeper et al., 2015). Determination of heterozygosity percent, UPGMA clustering, and PCoA analyses of the accessions were carried out in TASSEL v5.2.64 (Bradbury et al., 2007). To elucidate the genetic structure among cultivars, a Bayesian clustering algorithm approach was employed via the admixture model with correlated allele frequencies in STRUCTURE v2.3.4 (Pritchard et al., 2000). Before this, the SNP containing (.vcf) file was converted to the corresponding software input file formats by PGD Spider v2.1.1.5 (Lischer and Excoffier, 2012). A total of 10 independent runs were evaluated for 1 to 5 genetic subgroups, which were assumed to infer the optimum number of populations (K) present within the accessions. For each run, a burn-in period of 50000 and a Markov Chain Monte Carlo replication of 100000 were set. Most probable K (the number of subpopulations) was determined with the ΔK statistics via Structure Harvester v0.6.94 (Earl, 2012). The bar plots of population structure were generated with genotype labels and ordered by Q-value.

Functional effects of SNPs on genome

Identified high confidence SNPs among Darjeeling tea cultivars were mapped to the chromosome level reference genome of tea (Xia et al., 2020) to ascertain the putative candidate genes residing at their close vicinity. Associated SNPs effects on the genome were predicted using SnpEff v4.3 (Cingolani et al., 2012) based on a locally created database with the gene annotation (gff3) file retrieved from TPIA (Xia et al., 2019). Transcripts encompassing the variants within an exon, intron, splice site, and 5 Kbp up- or down-stream positions were recorded along with their description. The occurrence of SNP and their predicted effects were subsequently categorized as high (disruptive impact on the encoded protein), moderate (non-synonymous substitution), low (synonymous substitution), and modifier (impact on noncoding regions) impacts. Significant SNP associated genes were grouped according to their functions based on the KOG annotation database.

Marker trait association

Collectively 21 tea specific traits included in the present investigation as reported agronomic characters (yield, flavor, and stress tolerance) (Bhuyan et al., 2012; Singh, 1992) and our earlier findings on rigorous phenotyping of the cultivars (biochemical traits, Hazra et al. (2020a); stomatal characters, unpublished), were considered for possible marker-trait association studies. Accurately filtered 54,206 SNPs from 23 genotypes were considered for the GWAS. Genome-wide marker-trait association (MTA) analyses were conducted using

TASSEL v5.2.64 software (Bradbury et al., 2007) with the GLM protocol keeping a significant threshold for the association at $p < 0.05$ and $p < 0.001$. Furthermore, the Bonferroni equation derived p -value has been fixed as a threshold for the significance of putative associations between the SNPs and corresponding traits of interest. For the model fitting test of each trait, quantile-quantile (Q-Q) plots were generated showing the distribution of expected and observed p -values from the association test output. The Manhattan and QQ plots were represented using the ‘qqman’ R package (Turner, 2014) displaying the significant trait-associated markers.

Result

Genotyping

The present study acquired a total of 69,540,256 raw reads through the ddRAD based paired-end (150×2) sequencing on an Illumina Hiseq X10 platform. This output spanned an average of 3.02 M reads and 43.45 % GC content per accession with the highest (5.73 M) and lowest (1.46 M) read observed for the cultivars Badamtam-15/263 and Teesta Valley-1 respectively (Table S1). Pre-processed reads with RAD tag at both loci (98.44 ± 0.5 in percent) were mapped to the tea plant (*Camellia sinensis* var. *sinensis*) reference genome and resulted in a mean of 90.5% alignment rate (Table S1). Initially, 2,120,435 SNPs were identified through the variant calling pipeline, which on screening with 10X read depth and Q30 quality score cut-off, retained a total high quality 2,24,966 SNPs (Table S2).

Nucleotide substitution pattern of the variants included 1,65,567 transitions (Ts) and 59,399 transversions (Tv), with the transitions to transversions ratio of 2.79. The incidences of various substitutions were 36.9% C/T, 36.7% A/G, 6.2% A/T, 7.3% A/C, 7.4% G/T, and 5.5% C/G (Figure 1A). Tajima’s D and Minor allele frequency distributions of the polymorphic sites indicated that a predominant amount of positive Tajima’s D value and abundant sites with low MAF were present (Figure 1B-C). SNPs mapped to the chromosomes ranged from an average 82.8 (Chr3) to 121.6 (Chr1) per megabase (Table 1, Figure 1D). Variant sites residing in 15 chromosomes were further filtered with stringent parameters (minor allele frequency ≥ 0.05 , missing data within a locus < 0.3) for high confidence marker-trait association analyses, which returned a total of 54,206 SNPs.

Varietal diversity and genetic structure

Genome-wide SNPs were utilized to clarify the genetic structure among the studied elite cultivars of Darjeeling tea. The optimal number of subgroups (K) determined by STRUCTURE software run and subsequent Structure Harvester evaluation suggested peak ΔK value at $K = 2$ followed by $K = 3$ (Figure 2A). Bar plots of structure analyses considering $K = 2$, grouped the cultivars like MB-6, P-312, B-777, AV-2, and S-1 in a single subpopulation, TS-569 and RR-17/144 to be admixed and the remaining accessions classified to another set (Figure 2B). When ancestral components are assumed $K = 3$, the formerly mentioned group were intact as a single population, however, the larger part of accessions displayed admixture between two other subgroups at varying degrees. In the same output, cultivars like TS-569 showed a membership of all three subgroups (Figure 2C).

UPGMA dendrogram analysis has clustered the accessions in two distinct subgroups (Figure 2D). One of them was grouped by MB-6, B-777, P-312, AV-2, S-1, and TS-569 being separated from the main subgroup. In other subgroups, few cultivars strongly clustered with each other suggesting their genetic similarity, such as T-135 and B/5/63 belonged to the same branch. Likewise, two other cultivar pairs - B/6/62, SP/4/5, and T-78, B/15/263 shared a single subcluster of each. Results of the principal coordinate analyses were strongly corroborating to the genetic structure and dendrogram analyses (Figure 2E). Particular coordinate held by an accession and occurrence of others in the vicinity reflected the major subgroups as revealed by $K = 2$ and $k = 3$ in structure results as well as sub-clustering patterns of each cultivar as observed in the tree. Percentage of heterozygosity within the accessions indicated the lowest amount of heterozygous sites in TtV-1 (14.54) and B/6/62 (14.76), while the maximum value occurred in B/15/263 (27.94), AV-2 (26.75), T-78 (26.5), and HV39 (25.8). Heterozygosity of the remaining cultivars are almost similar and lies between 17-22% (Figure 2F).

SNP effects on genome

Prediction of 54,206 well-confidenced SNPs putative effect on tea genome exhibited most of them having modifier effect (96.06%), followed by moderate effect (2.1%). Least occurrence was observed in case of high impact (0.22%) *i.e.* with disruptive impact on the encoded protein. SNPs with modifier impact were frequently observed in non-coding genomic regions, which are intergenic positions (75.96%) followed by intron variants (8.33%) and within 5kb downstream (6.86%) or upstream (4.22%) of the annotated genes. The present

study revealed a total of 3177 genes in the tea genome to be impacted by SNPs at varying degrees. According to the functional impact classification of the SNPs in the exon region, missense type was the most prevalent one (59.79%) and a little expense (1.79%) of nonsense mutations were observed. A considerable amount of variants occurred in acceptor or donor sites of splicing regions too (Figure 3, Table S3).

A total of 95 genes were identified to harbor variant sites that have a potential impact on the gene function by altering their protein product and thereby probable phenotype modification of the source plant (Table S4). Most of these genes were with at least one SNP except for CSS0030092 which had four and CSS0004975, CSS0008917, CSS0011515, CSS0017046, CSS0019151, CSS0029404 with three in each, and 28 genes having two SNPs. The KOG class prediction ensued that a major fraction of SNPs impacted genes are involved in posttranslational modification, signal transduction mechanisms, RNA processing, and transcript modifications, etc (Figure 3). However, 46.74% of the candidate genes found no hit to the KOG database, and 16.18% of them were poorly characterized (*i.e.* General function prediction only or unknown function) (Table S5). Predicted changes on encoded amino acids from variant coding sites led to various modes of consequences at the protein level. These include nonpolar to polar (Ala-Thr, Gly-Arg, Pro-Arg, etc.), negatively charged to positively charged or non-charged (Glu-Lys, Asp-Asn, etc.) and aromatic to aliphatic (Phe-Leu) residue substitutions (Table S6) as a major fraction of occurrences.

Genome-wide association studies

The core set of high quality ($MAF \geq 0.05$, missing data < 0.3) SNPs were chosen for association analyses with 21 agronomic and health benefit traits that are specific to tea. The probability of association between each trait and putatively associated variant sites was depicted using Manhattan and QQ plots (Figure 4a, b). A certain amount of loci were observed to show covariance with the phenotypic traits at varying probability threshold ($p < 0.05$ and $p < 0.001$) (Table S7). However, to eliminate the low confidence associations and reduce false-positive rates, a stringent screening was performed with the manual examination of expected and observed probabilistic QQ plots followed by the selection of putatively associated SNPs by a Bonferroni correction threshold ($p < 1.845 \times 10^{-5}$). Accordingly, a total of 57 markers were identified to be significantly associated with eight phenotypic traits (Table 2). These trait-associated markers were distributed across the 14 out of 15 chromosomes of tea. Most of the significant associations observed in case of various secondary metabolites,

such as EGCG and flavonoids. Mapping of the SNPs on the genome resulted in a major amount of the markers localized on intergenic regions. However, a few were observed to occur in the downstream, upstream, or intron portions of the gene as well. One flavonoid related SNP was observed to cause a missense type of mutation in the CSS0000649 gene. KOG class prediction of the trait-associated SNP impacted genes resulted in signal transduction related genes is the most predominant followed by the post-transcriptional and post-translational type.

Discussion

With the advent of molecular genetic approaches, efficient genotyping with high throughput markers has led to an in-depth understanding of crop genetic resources, their management, and utilization in breeding programs (Basak et al., 2019; Sudan et al., 2019). Development of SNP markers along with the determination of their genomic position assisted the trait association analyses in plants and their diversity at an individual, community as well as species level (Guajardo et al., 2020; Huq et al., 2016; Luo et al., 2019; Mammadov et al., 2012). In the present study, ddRAD based genotyping by sequencing approach was adopted for genome-wide SNP discovery and their diversity assessment in 23 Darjeeling tea cultivars. These accessions represent contrasting genetic backgrounds, most of which are cultivated in the terrain of Darjeeling hills for their preferential quality traits. The double-digest strategy and deep coverage paired-end sequencing approach have recommended toward accomplishing greater quality SNPs, lower amounts of missing data, and precise mapping onto the reference genome (Guajardo et al., 2020; Shirasawa et al., 2016). Based on the 54K high-quality SNPs developed in this study and mapped across 15 chromosomes of tea, the genetic relatedness of the Darjeeling tea cultivars has been established. The degree of relatedness inferred through Bayesian clustering, UPGMA method, and PCoA analyses conform to each other. However, the clustering does not result in distinct lineages since the accessions carry an enormous gene pool of highly variable cultivars of Indian hybrid tea. Most of the commercial accessions considered in the present study belong to a group of ‘mixed’ ancestry as revealed by the population genetic structure of Indian tea (Raina et al., 2012). Optimum number of K obtained here for K = 2, still, three subpopulations have been assumed to concealed within the accessions as three separate origins of domesticated tea lineages exists (Meegahakumbura et al., 2018; Yamashita et al., 2019). In STRUCTURE analyses of selected Darjeeling tea cultivars, homogeneity was observed in a major part when

two subpopulations were assumed. However, the same members had shown admixture between two different groups at varying degrees when $K = 3$ was considered. Both analyses revealed the genetic background of TS-569 cultivar to be admixed between two separate ancestries, which is following its true biparental origin ($AV-2 \times T-78$) (Hazra et al., 2020b). The high level of heterozygosity indices (15-30%) of the studied accessions observed in the present study corroborates their ‘hybrid’ structures too. Following ddRAD based study on Chinese tea, Yang et al. (2016) also opined that due to self-incompatibility, domestication via hybridization, and climatic selection, cultivars might have resulted in broader genetic variation. A large number of high-resolution SNPs spanning tea genome thereby discovered exclusively from Darjeeling tea, which not only can distinguish the accessions with the desirable trait, but also support for understanding the effect of these variations through functional genomics aspects. Since genomic variations in the worldwide germplasm of tea is still under discovery, therefore, variants identified from this study would serve the reference collection.

Non-synonymous single nucleotide variations with the potential to alter the associated gene functions are valued in molecular breeding studies for their profound role in trait relations and marker-assisted selection (Cao et al., 2016; Guajardo et al., 2020; Shirasawa et al., 2013). Although a major fraction of SNPs discovered in this study, found to be located in the intergenic region, a substantial amount occurred in the genic region too. Among which SNPs detected in the non-coding regulatory regions such as upstream, downstream, UTR, or even intron regions might affect the gene function at the transcriptional and/or translational level. These changes along with the exon level variants should be validated on a case-by-case basis because not all SNPs in the genic or associated region are functionally impactful (Cingolani et al., 2012). The present study reveals that the occurrences of SNPs imposing modifier effects were most abundant. Similar observations were reported earlier, experimented with crops like soybean (Ramakrishna et al., 2018), pear (Montanari et al., 2019), and *Prunus* (Guajardo et al., 2020). Here, based on functionality, only a set of 95 genes among the total 3177 SNP impacted genes in the tea genome, could be affected highly. Because SNPs in these genes predicted to cause either stop codon gain/loss or frame-shift, missense, splice donor, and acceptor variants (Table S4). Functional annotations of high impact SNP containing genes disclosed a large proportion of them with unknown or poorly characterized function. However, the rest of them annotated to perform functions including signal transduction, secondary metabolite biosynthesis, transcriptional and translational regulation,

cell cycle control, and cellular transport. Nevertheless, SNPs residing in candidate genes need further functional genomics validation to examine their impact on the consequent phenotypic effect.

Genome-wide association study has become a prospective approach for the genetic dissection of complex traits and it has successfully been adopted in various crops (Jaiswal et al., 2019; Luo et al., 2019; Raggi et al., 2019; Sudan et al., 2019). However, complex phenotypic traits linked molecular markers are still inadequate in tea to be utilized for marker-assisted selection (Hazra et al., 2019). To identify whether any nucleotide substitutions observed in the scope of present work underlying a phenotyping change, GWAS has been considered. It reveals trait-associated markers are distributed in most of the chromosomes revealing the significant role of each of the chromosomes in controlling the studied traits. Although, a large number of variants showed trait relations at a low confidence level ($p < 0.05$), only 1% or less retained at a more restricted parameter ($p < 0.001$) for most of the traits. In this study, a low sample size may stand as a limiting factor instigating a reduced significance level for the individual marker-trait association. Finally, the Bonferroni correction endorsed hardly a few loci (57 out of 54,206) to cross the threshold. The Bonferroni correction method, which estimates a threshold p -value to control the false positive rate, is based on the assumption that every genetic variant tested is independent of the rest (Kaler and Purcell, 2019). However, this is also considered as the most conservative method and many important loci may not qualify the stringent criterion of significance test (Wang et al., 2016). Hence, each of the markers identified can be utilized later on, for a larger population (especially from a mapping population) to check their true associations and heritability.

Since most agronomic traits are polygenic, background genetic variation of such complex traits exists in multiple loci. Genic and intergenic SNPs collectively perhaps underlie the traits and thus, understanding their functional implications is important prior to employ them in breeding studies. Exploration of the SNP surrounding genomic regions allowed the identification of several candidate genes that might have a vital role in controlling the traits. One missense variant in CSS0000649 and another variation within the upstream region of CSS0032954 exhibited significant association with flavonoid accumulation of fresh tea leaves. Functional prediction revealed the involvement of these genes in the signal transduction mechanism. Likewise, functionally relevant genic SNPs, reported in this study, might be validated and surveyed on an extensive worldwide collection of wild and cultivated

genotypes of tea, so that their genealogy through domestication or artificial selection can be traced.

Acknowledgment

Authors are thankful to the respective authority of Darjeeling Tea Research and Development Center, Kurseong, India for providing necessary permissions regarding fieldwork and to use elite accessions of Darjeeling tea for research purposes.

Author Contribution

Conceptualization: AH and SD, fieldwork: AH and RK, laboratory experiments: AH, data interpretation: AH, CS and SD, manuscript preparation: AH and SD

References

- Afgan, E., Baker, D., Batut, B., Van Den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research* 46, W537-W544.
- Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
- Basak, M., Uzun, B., Yol, E., 2019. Genetic diversity and population structure of the Mediterranean sesame core collection with use of genome-wide SNPs developed by double digest RAD-Seq. *PloS one* 14, e0223757.
- Bhuyan, L., Senapati, K.K., Saikia, P., Hazarika, M., 2012. Characterization of volatile flavour constituents of orthodox black tea of twenty nine Tocklai released cultivars for Darjeeling. *Two Bud* 59, 112-118.
- Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114-2120.
- Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., Buckler, E.S., 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633-2635.
- Cao, K., Zhou, Z., Wang, Q., Guo, J., Zhao, P., Zhu, G., Fang, W., Chen, C., Wang, X., Wang, X., 2016. Genome-wide association study of 12 agronomic traits in peach. *Nature Communications* 7, 1-10.
- Chen, J., Wang, B., Zhang, Y., Yue, X., Li, Z., Liu, K., 2017. High-density ddRAD linkage and yield-related QTL mapping delimits a chromosomal region responsible for oil content in rapeseed (*Brassica napus* L.). *Breeding science*, 16116.

- Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6, 80-92.
- Das, K., 2006. International protection of India's geographical indications with special reference to “Darjeeling” tea. *The Journal of World Intellectual Property* 9, 459-495.
- Davik, J., Sargent, D.J., Brurberg, M.B., Lien, S., Kent, M., Alsheikh, M., 2015. A ddRAD based linkage map of the cultivated strawberry, *Fragaria xananassa*. *PLoS One* 10, e0137746.
- Dereeper, A., Homa, F., Andres, G., Sempere, G., Sarah, G., Hueber, Y., Dufayard, J.-F., Ruiz, M., 2015. SNIPlay3: a web-based application for exploration and large scale analyses of genomic variations. *Nucleic Acids Research* 43, W295-W300.
- Earl, D.A., 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation genetics resources* 4, 359-361.
- Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* 6, e19379.
- Firmani, P., De Luca, S., Bucci, R., Marini, F., Biancolillo, A., 2019. Near infrared (NIR) spectroscopy-based classification for the authentication of Darjeeling black tea. *Food Control* 100, 292-299.
- Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.
- Guajardo, V., Solís, S., Almada, R., Saski, C., Gasic, K., Moreno, M.Á., 2020. Genome-wide SNP identification in *Prunus* rootstocks germplasm collections using Genotyping-by-Sequencing: phylogenetic analysis, distribution of SNPs and prediction of their effect on gene function. *Scientific reports* 10, 1-14.
- Hazra, A., Dasgupta, N., Kumar, R., Sengupta, C., Das, S., 2020a. miRNA precursor-derived SSR marker-mediated genotyping of tea (*Camellia sinensis*) cultivars with varying health benefit traits. *Plant Breeding*.
- Hazra, A., Dasgupta, N., Sengupta, C., Bera, B., Das, S., 2019. Tea: A Worthwhile, Popular Beverage Crop Since Time Immemorial, *Agronomic Crops*. Springer, pp. 507-531.
- Hazra, A., Dasgupta, N., Sengupta, C., Das, S., 2018a. Next generation crop improvement program: Progress and prospect in tea (*Camellia sinensis* (L.) O. Kuntze). *Annals of Agrarian Science* 16, 128-135.
- Hazra, A., Dasgupta, N., Sengupta, C., Kumar, R., Das, S., 2018b. On some biochemical physiognomies of two common Darjeeling tea cultivars in relation to blister blight disease. *Archives of Phytopathology and Plant Protection* 51, 915-926.

- Hazra, A., Saha, S., Kumar, R., Chaubey, M., Sengupta, C., Das, S., 2020b. Efficacy of several QTL and miRNA-SSR markers towards health benefit traits in an elite Darjeeling tea cultivar. *Journal of Microbiology, Biotechnology and Food Sciences* 9, 700-704.
- Hirakawa, H., Shirasawa, K., Ohyama, A., Fukuoka, H., Aoki, K., Rothan, C., Sato, S., Isobe, S., Tabata, S., 2013. Genome-wide SNP genotyping to infer the effects on gene functions in tomato. *DNA research* 20, 221-233.
- Huq, M.A., Akter, S., Nou, I.S., Kim, H.T., Jung, Y.J., Kang, K.K., 2016. Identification of functional SNPs in genes and their effects on plant phenotypes. *Journal of Plant Biotechnology* 43, 1-11.
- Jaiswal, V., Gupta, S., Gahlaut, V., Muthamilarasan, M., Bandyopadhyay, T., Ramchiary, N., Prasad, M., 2019. Genome-wide association study of major agronomic traits in foxtail millet (*Setaria italica* L.) using ddRAD sequencing. *Scientific reports* 9, 1-11.
- Kaler, A.S., Purcell, L.C., 2019. Estimation of a significance threshold for genome-wide association studies. *BMC genomics* 20, 618.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. *Genome research* 19, 1639-1645.
- Laila, R., Park, J.-I., Robin, A.H.K., Natarajan, S., Vijayakumar, H., Shirasawa, K., Isobe, S., Kim, H.-T., Nou, I.-S., 2019. Mapping of a novel clubroot resistance QTL using ddRAD-seq in Chinese cabbage (*Brassica rapa* L.). *BMC plant biology* 19, 13.
- Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10, R25.
- Lee, J.-H., Natarajan, S., Biswas, M.K., Shirasawa, K., Isobe, S., Kim, H.-T., Park, J.-I., Seong, C.-N., Nou, I.-S., 2018. SNP discovery of Korean short day onion inbred lines using double digest restriction site-associated DNA sequencing. *PloS one* 13, e0201229.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078-2079.
- Lischer, H.E., Excoffier, L., 2012. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28, 298-299.
- Luo, Z., Tomasi, P., Fahlgren, N., Abdel-Haleem, H., 2019. Genome-wide association study (GWAS) of leaf cuticular wax components in *Camelina sativa* identifies genetic loci related to intracellular wax transport. *BMC plant biology* 19, 187.
- Ma, B., Liao, L., Peng, Q., Fang, T., Zhou, H., Korban, S.S., Han, Y., 2017. Reduced representation genome sequencing reveals patterns of genetic diversity and selection in apple. *Journal of integrative plant biology* 59, 190-204.

- Mammadov, J., Aggarwal, R., Buyyarapu, R., Kumpatla, S., 2012. SNP markers and their impact on plant breeding. *International journal of plant genomics* 2012.
- Meegahakumbura, M.K., Wambulwa, M.C., Li, M.-M., Thapa, K.K., Sun, Y.-S., Möller, M., Xu, J.-C., Yang, J.-B., Liu, J., Liu, B.-Y., 2018. Domestication origin and breeding history of the tea plant (*Camellia sinensis*) in China and India based on nuclear microsatellites and cpDNA sequence data. *Frontiers in plant science* 8, 2270.
- Montanari, S., Bianco, L., Allen, B.J., Martínez-García, P.J., Bassil, N.V., Postman, J., Knäbel, M., Kitson, B., Deng, C.H., Chagné, D., 2019. Development of a highly efficient Axiom™ 70 K SNP array for *Pyrus* and evaluation for high-density mapping and germplasm characterization. *BMC genomics* 20, 331.
- Mukhopadhyay, M., Mondal, T.K., Chand, P.K., 2016. Biotechnological advances in tea (*Camellia sinensis* [L.] O. Kuntze): a review. *Plant cell reports* 35, 255-287.
- Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one* 7, e37135.
- Phukan, M., Savapondit, D., Hazra, A., Das, S., Pramanik, P., 2018. Algorithmic derivation of CO₂ assimilation based on some physiological parameters of tea bushes in North-East India. *Ecological Indicators* 91, 77-83.
- Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. *Genetics* 155, 945-959.
- Raggi, L., Caproni, L., Carboni, A., Negri, V., 2019. Genome-wide association study reveals candidate genes for flowering time variation in common bean (*Phaseolus vulgaris* L.). *Frontiers in plant science* 10, 962.
- Raina, S.N., Ahuja, P., Sharma, R., Das, S., Bhardwaj, P., Negi, R., Sharma, V., Singh, S., Sud, R., Kalia, R., 2012. Genetic structure and diversity of India hybrid tea. *Genetic resources and crop evolution* 59, 1527-1541.
- Ramakrishna, G., Kaur, P., Nigam, D., Chaduvula, P.K., Yadav, S., Talukdar, A., Singh, N.K., Gaikwad, K., 2018. Genome-wide identification and characterization of InDels and SNPs in *Glycine max* and *Glycine soja* for contrasting seed permeability traits. *BMC plant biology* 18, 141.
- Ray, S., Satya, P., 2014. Next generation sequencing technologies for next generation plant breeding. *Frontiers in plant science* 5, 367.
- Roy, S.C., Moitra, K., De Sarker, D., 2017. Assessment of genetic diversity among four orchids based on ddRAD sequencing data for conservation purposes. *Physiology and molecular biology of plants* 23, 169-183.

- Sasai, R., Tabuchi, H., Shirasawa, K., Kishimoto, K., Sato, S., Okada, Y., Kuramoto, A., Kobayashi, A., Isobe, S., Tahara, M., 2019. Development of molecular markers associated with resistance to *Meloidogyne incognita* by performing quantitative trait locus analysis and genome-wide association study in sweetpotato. *DNA Research* 26, 399-409.
- Scheben, A., Batley, J., Edwards, D., 2017. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant biotechnology journal* 15, 149-161.
- Shirasawa, K., Fukuoka, H., Matsunaga, H., Kobayashi, Y., Kobayashi, I., Hirakawa, H., Isobe, S., Tabata, S., 2013. Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato. *DNA research* 20, 593-603.
- Shirasawa, K., Hirakawa, H., Isobe, S., 2016. Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato. *DNA research* 23, 145-153.
- Singh, G., 2019. Can climate-resilient clones save Darjeeling tea?, Mongabay.
- Singh, I., 1992. Tea cultivars for Darjeeling. Report published by Tea Research Association, Nagrakata Sub Station, Nagrakata, Jalpaiguri, West Bengal, India.
- Sudan, J., Singh, R., Sharma, S., Salgotra, R.K., Sharma, V., Singh, G., Sharma, I., Sharma, S., Gupta, S.K., Zargar, S.M., 2019. ddRAD sequencing-based identification of inter-genepool SNPs and association analysis in *Brassica juncea*. *BMC plant biology* 19, 1-15.
- Turner, S.D., 2014. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *Biorxiv*, 005165.
- Wang, X., Gui, S., Pan, L., Hu, J., Ding, Y., 2016. Development and characterization of polymorphic microRNA-based microsatellite markers in *Nelumbo nucifera* (Nelumbonaceae). *Applications in plant sciences* 4, 1500091.
- Xia, E., Tong, W., Hou, Y., An, Y., Chen, L., Wu, Q., Liu, Y., Yu, J., Li, F., Li, R., 2020. The Reference Genome of Tea Plant and Resequencing of 81 Diverse Accessions Provide Insights into Its Genome Evolution and Adaptation. *Molecular plant* 13, 1013-1026.
- Xia, E.H., Li, F.D., Tong, W., Li, P.H., Wu, Q., Zhao, H.J., Ge, R.H., Li, R.P., Li, Y.Y., Zhang, Z.Z., 2019. Tea Plant Information Archive: a comprehensive genomics and bioinformatics platform for tea plant. *Plant biotechnology journal* 17, 1938-1953.
- Yamashita, H., Katai, H., Kawaguchi, L., Nagano, A.J., Nakamura, Y., Morita, A., Ikka, T., 2019. Analyses of single nucleotide polymorphisms identified by ddRAD-seq reveal genetic structure of tea germplasm and Japanese landraces for tea breeding. *PloS one* 14, e0220981.

Yang, H., Wei, C.-L., Liu, H.-W., Wu, J.-L., Li, Z.-G., Zhang, L., Jian, J.-B., Li, Y.-Y., Tai, Y.-L., Zhang, J., 2016. Genetic divergence between *Camellia sinensis* and its wild relatives revealed via genome-wide SNPs from RAD sequencing. PLoS One 11, e0151424.

Figure and table legends

Figure 1 – **Characteristics of genome-wide SNPs identified.** A – Transition and transversion fractions, B – Tajima's D distribution of identified variant sites, C – minor allele frequency distribution of variant sites with 10X read depth, D – Distribution of SNPs along the 15 chromosomes of tea genome (*Camellia sinensis* var. *sinensis*). SNP density was plotted in 1 Mbp sliding window using Circos. Heatmap and histogram tracks from outside to inside represents distribution of initial high depth (10X) 2,24,966 SNPs and subsequently filtered (MAF < 0.05, minimum SNP call in a locus > 0.7) 54,206 SNPs respectively.

Figure 2 – **Genetic relatedness and heterozygosity levels of 23 Darjeeling tea cultivars.** A – Optimal number of subpopulations estimated by plotting ΔK against each K of STRUCTURE analyses, B – Inferred genetic structure at K = 2 and C - K = 3, D – UPGMA dendrogram and E – Principal coordinate analyses with the identified SNP markers, F - Heterozygosity percentages of the tea accessions.

Figure 3 – Classification of detected SNPs according to their effect and distribution across various genomic regions (Pie charts). The bar graph represents the KOG annotation-based classification of genes that harboring variant sites with potential effects.

Figure 4a-b - Manhattan and quantile-quantile (Q-Q) plot based on the GWAS results for eight selected traits.

Table 1 - Distribution of identified SNPs mapped to each chromosome.

Table 2 - Summary of trait-associated markers that accomplished the Bonferroni correction threshold.

Figure S1 – Agarose gel electrophoresis of isolated genomic DNA from 23 accessions (M – 1kb ladder)

Table S1 – List of accessions selected for the study, statistics of DNA, sequence quality, and alignment rates.

Table S2 – Summary of generated reads and variant calling

Table S3 – Records of SNP effects and putative impacts on the tea genome predicted by the SNPeff program.

Table S4 – Variant containing genes and summary of predicted effects.

Table S5 – Classification of SNP harboring genes based on KOG annotations

Table S6 - Substitution statistics of amino acid residues caused by SNPs.

Table S7 – Amount of trait-associated loci obtained for each trait with varying levels of significance.

Table 1 – Distribution of identified SNPs mapped to each chromosome.

Chromosome	Length in bp	No. of genes residing	Detected SNPs	Mean SNPs per Mb
Chr1	2.23E+08	3850	27106	121.6
Chr2	2.13E+08	3814	22660	106.3
Chr3	1.87E+08	3258	15519	82.8
Chr4	1.96E+08	3314	19684	100.3
Chr5	1.95E+08	2843	18283	93.8
Chr6	1.81E+08	3332	16745	92.4
Chr7	1.87E+08	3301	18884	100.9
Chr8	1.63E+08	2237	18547	113.7
Chr9	1.66E+08	2737	17884	108.0
Chr10	1.67E+08	2492	15982	95.6
Chr11	1.24E+08	2374	13808	111.6
Chr12	1.63E+08	2332	15626	96.0
Chr13	1.36E+08	2345	13209	97.0
Chr14	1.34E+08	2367	13771	102.9
Chr15	1.19E+08	2009	11929	100.2
Unassigned Contigs	2.99E+08	7920	20971	70.1

Table 2 – Summary of trait associated markers that accomplished the Bonferroni correction threshold.

Trait	Marker	Chromosome	Position	Variant	Minor Allele Frequency	marker R ²	p	Genomic region	Associated genes
EGCG	C11_107011137	11	107011137	A/G	0.118	0.508	1.4E-06	Intergenic region	CSS0006764-CSS0021815
	C12_8344293	12	8344293	G/T	0.441	0.531	1.6E-06	Upstream	CSS0021925
	C2_23454393	2	23454393	G/A	0.125	0.549	3.6E-07	Intergenic region	CSS0041829-CSS0003610
	C3_42104228	3	42104228	G/T	0.342	0.584	1.3E-06	Intron	CSS0018073
	C3_96267049	3	96267049	A/C	0.125	0.513	1.4E-06	Intron	CSS0025631
	C4_137059167	4	137059167	G/A	0.361	0.617	1.3E-06	Intergenic region	CSS0043500-CSS0003582
	C5_40077092	5	40077092	A/T	0.176	0.793	1.4E-07	Intergenic region	CSS0001517-CSS0014046
	C5_77389054	5	77389054	G/C	0.065	0.555	4E-07	Intergenic region	CSS0037593-CSS0007042
	C6_133123050	6	133123050	A/G	0.158	0.494	1.7E-06	Downstream	CSS0005970
	C6_150191592	6	150191592	C/T	0.306	0.705	5.3E-07	Intergenic region	CSS0022296-CSS0027426
Flavonoids	C6_150191604	6	150191604	G/A	0.333	0.705	5.3E-07	Intergenic region	CSS0022296-CSS0027426
	C6_150191625	6	150191625	A/C	0.361	0.706	5E-07	Intergenic region	CSS0022296-CSS0027426
	C12_146118907	12	146118907	G/T	0.056	0.502	1.4E-06	Intergenic region	CSS0035382-CSS0012355
	C12_96808427	12	96808427	G/A	0.500	0.429	1.6E-06	Intergenic region	CSS0033122-CSS0030288
	C12_96808473	12	96808473	C/T	0.313	0.429	1.6E-06	Intergenic region	CSS0033122-CSS0030288
	C15_106968540	15	106968540	C/G	0.143	0.561	1.6E-06	Exon (mis-sense)	CSS0000649
	C3_132783655	3	132783655	G/T	0.469	0.820	7.5E-07	Intergenic region	CSS0016870-CSS0028807
	C6_60012823	6	60012823	G/A	0.071	0.501	1.4E-06	Upstream	CSS0032954
	C7_92092737	7	92092737	C/T	0.438	0.738	8.8E-08	Intergenic region	CSS0048867-CSS0040559
	C8_40520195	8	40520195	G/T	0.225	0.574	9E-07	Exon (Synonymous)	CSS0048533
Flavour	C1_158342510	1	158342510	C/T	0.375	0.431	1.8E-06	3' UTR	CSS0046973
	C11_36853670	11	36853670	G/A	0.441	0.296	7.2E-07	Downstream	CSS0038387
	C11_36853672	11	36853672	T/C	0.441	0.296	7.2E-07	Downstream	CSS0038387
	C3_178449714	3	178449714	A/C	0.364	0.523	1.3E-06	Downstream ; Intron	CSS0003232 ; CSS0017256
	C4_23408141	4	23408141	A/G	0.306	0.476	8.3E-07	Intergenic region	CSS0010887-CSS0000279
	C9_129576673	9	129576673	C/T	0.281	0.258	1.4E-06	Intergenic region	CSS0045989-CSS0029293
	C9_163100327	9	163100327	G/A	0.281	0.212	1.1E-06	Intergenic region	CSS0017572-CSS0036456
	C9_163100375	9	163100375	C/A	0.281	0.212	1.1E-06	Intergenic region	CSS0017572-CSS0036456
	C4_161227403	4	161227403	A/G	0.472	0.449	2E-06	Intergenic region	CSS0043770-CSS0034804
	C4_161227418	4	161227418	C/G	0.472	0.449	2E-06	Intergenic region	CSS0043770-CSS0034804
Phenolics	C4_161227493	4	161227493	A/G	0.472	0.449	2E-06	Intergenic region	CSS0043770-CSS0034804
	C1_106410162	1	106410162	C/T	0.395	0.776	3.3E-07	Intergenic region	CSS0004784-CSS0014147
	C1_106410245	1	106410245	G/C	0.395	0.776	3.3E-07	Intergenic region	CSS0004784-CSS0014147
	C10_105898824	10	105898824	G/A	0.211	0.498	1.8E-06	Intergenic region	CSS0005939-CSS0038872
	C10_17378101	10	17378101	T/C	0.342	0.694	5.7E-07	Intergenic region	CSS0005636-CSS0013814
	C11_11657035	11	11657035	C/T	0.325	0.681	4.6E-07	Intergenic region	CSS0000131-CSS0039854
	C11_11657088	11	11657088	T/C	0.325	0.681	4.6E-07	Intergenic region	CSS0000131-CSS0039854
	C11_4149052	11	4149052	G/A	0.381	0.574	1.5E-06	Downstream ; Intron	CSS0023670 ; CSS0003583
	C5_12705703	5	12705703	G/A	0.225	0.660	1E-06	Intergenic region	CSS0035557-CSS0015046
	C3_122306599	3	122306599	C/T	0.059	0.591	3.6E-06	Intergenic region	CSS0024042-CSS0016916
Stomatal Index	C4_13092427	4	13092427	C/T	0.059	0.515	6.5E-06	Intergenic region	CSS0033093-CSS0019787
	C5_20023429	5	20023429	C/T	0.079	0.646	6.9E-06	Intergenic region	CSS0036130-CSS0005128
	C6_70825915	6	70825915	A/G	0.053	0.601	2.6E-06	Intergenic region	CSS0006721-CSS0001755
	C11_18591436	11	18591436	C/A	0.150	0.635	4.3E-06	Downstream	CSS0017212
	C13_13274794	13	13274794	C/A	0.150	0.565	7.6E-06	Intergenic region	CSS0019314-CSS0039847
	C1_136784684	1	136784684	T/G	0.353	0.445	3.2E-06	Intergenic region	CSS0038385-CSS0033601
	C13_46382138	13	46382138	A/G	0.278	0.435	5.9E-06	Intergenic region	CSS0045066-CSS0031411
	C13_46382173	13	46382173	G/C	0.278	0.435	5.9E-06	Intergenic region	CSS0045066-CSS0031411
	C2_108441184	2	108441184	C/T	0.167	0.379	3.6E-06	Intergenic region	CSS0020396-CSS0036838
	C2_70418684	2	70418684	C/T	0.344	0.538	1.4E-06	Intergenic region	CSS0006855-CSS0032238
Yield	C11_36571066	11	36571066	G/A/C	0.219	0.584	6.9E-07	Intergenic region	CSS0008620-CSS0029192
	C12_38185130	12	38185130	C/T	0.316	0.775	4.2E-07	Intergenic region	CSS0018260-CSS0001772
	C12_80732309	12	80732309	A/G	0.294	0.760	1.7E-06	Intron	CSS0022785
	C15_71441068	15	71441068	T/A/G	0.375	0.789	5.2E-07	Intergenic region	CSS0030870-CSS0027046
	C3_84356131	3	84356131	G/C	0.059	0.412	1.9E-06	Intergenic region	CSS0028825-CSS0024256
	C4_74941084	4	74941084	G/A	0.114	0.636	2E-06	Intergenic region	CSS0001362-CSS0032765
	C4_74941144	4	74941144	T/C	0.114	0.636	2E-06	Intergenic region	CSS0001362-CSS0032765

Figure 1

Figure 3

Figure 4a

Figure 4b

Figure S1