DNA Research **20**, 593–603, (2013)
Advance Access publication on 31 July 2013

doi:10.1093/dnares/dst033

Genome-Wide Association Studies Using Single Nucleotide Polymorphism Markers Developed by Re-Sequencing of the Genomes of Cultivated Tomato

Kenta Shirasawa[1,*], Hiroyuki Fukuoka[2], Hiroshi Matsunaga[2], Yuhko Kobayashi[3], Issei Kobayashi[3], Hideki Hirakawa[1], Sachiko Isobe[1], and Satoshi Tabata[1]

Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan[1]; NARO Institute of Vegetable and Tea Sciences, 360 Kusawa, Ano, Tsu, Mie 514-2392, Japan[2] and Life Science Research Center, Mie University, 1577 Kurimamachiya, Tsu, Mie 514-8507, Japan[3]

*To whom correspondence should be addressed. Tel. +81 438-52-3935. Fax. +81 438-52-3934. E-mail: shirasaw@kazusa.or.jp

Edited by Dr Katsumi Isono
(Received 4 June 2013; accepted 5 July 2013)

Abstract

With the aim of understanding relationship between genetic and phenotypic variations in cultivated tomato, single nucleotide polymorphism (SNP) markers covering the whole genome of cultivated tomato were developed and genome-wide association studies (GWAS) were performed. The whole genomes of six tomato lines were sequenced with the ABI-5500xl SOLiD sequencer. Sequence reads covering ∼13.7× of the genome for each line were obtained, and mapped onto tomato reference genomes (SL2.40) to detect ∼1.5 million SNP candidates. Of the identified SNPs, 1.5% were considered to confer gene functions. In the subsequent Illumina GoldenGate assay for 1536 SNPs, 1293 SNPs were successfully genotyped, and 1248 showed polymorphisms among 663 tomato accessions. The whole-genome linkage disequilibrium (LD) analysis detected highly biased LD decays between euchromatic (58 kb) and heterochromatic regions (13.8 Mb). Subsequent GWAS identified SNPs that were significantly associated with agronomical traits, with SNP loci located near genes that were previously reported as candidates for these traits. This study demonstrates that attractive loci can be identified by performing GWAS with a large number of SNPs obtained from re-sequencing analysis.

Key words: genome-wide association studies; linkage disequilibrium; whole-genome re-sequencing; single nucleotide polymorphism; tomato

1. Introduction

Tomato (*Solanum lycopersicum*), which is considered to be an important crop, originated from South and Central America, and spread to the rest of the world with accompanying morphological diversification.[1] The Solanaceae family, to which tomato belongs, includes other important crop species, such as potato (*S. tuberosum*), eggplant (*S. melongena*), tobacco (*Nicotiana tabacum*), and pepper (*Capsicum annuum*). Comparative genomics within these various genera and species have greatly accelerated understanding of their genome evolution and the genetic mechanisms that confer phenotypic diversity to these species.[2] Furthermore, several interspecific genetic linkage maps have been constructed between cultivated tomato and its wild relatives (*S. chmielewskii*, *S. habrochaites*, *S. pennellii*, and *S. pimpinellifolium*).[3] These maps allow identification of the genes responsible for interspecific phenotypic variations, including disease resistance, fruit size and shape, and plant architecture.[3] However, few genetic studies have reported intraspecific variations due to its narrow genetic diversity.[3,4]

In the field of human and animal genomic and genetic studies, the availability of whole-genome sequence data has resulted in more rapid advances in re-sequencing analysis and genome-wide association studies (GWAS) than in classical genetics and quantitative trait locus (QTL) mapping.[5,6] In plants such as rice (*Oryza sativa*) and *Arabidopsis thaliana*, the initial plant species for which whole-genome sequences were available provided representative targets for such analysis.[7–10]

Tomato has also been used as a model plant in classical and molecular genetics,[11] due to autogamous diploidy ($2n = 2x = 24$) and a relatively compact genome (~950 Mb). Recently, the whole-genome sequence of tomato was published.[12] Furthermore, Hirakawa *et al.*[13] inferred the functions of 200 SNPs among the transcribed sequences of cultivated tomato lines by determining their positions in predicted genes on the tomato genome. These results are expected to accelerate the understanding of genetic mechanisms that confer phenotypic variations among tomato cultivars.

Massive parallel sequencing and genotyping methods have contributed to progress in genetics and genomics. Next-generation sequencers (NGSs), such as HiSeq2500 (Illumina), the GS FLX+ system (Roche), 5500xl SOLiD (Life Technologies), and Ion Proton (Life Technologies), have been employed for *de novo* assembly of genome sequences and re-sequencing analyses of genomes of several organisms.[14,15] In such re-sequencing analysis, sequence reads from the whole genome are mapped onto the reference genome to identify nucleotide variations, including single nucleotide polymorphisms (SNPs) and insertions/deletions (indels).[14] A large amount of nucleotide sequence data (up to Mb- or Gb-scale), redundantly covering the whole-genome sequence, can be obtained simultaneously by NGS technologies. This allows a huge number of the nucleotide variations to be identified cheaply and within a relatively short period of time. The identified SNPs can be used, for example, for polymorphic analysis of germplasm collections, which, in turn, allows genetic analyses such as QTL mapping, GWAS, and genomic selection.[16] Large-scale SNP genotyping is often performed with commercially available array-based platforms, such as Infinium (Illumina), GoldenGate (Illumina), and Axiom Genotyping Solution (Affymetrix).

Tomato accessions, so-called genetic resources, are stocked in several gene banks, including the Tomato Genetic Resource Center (TGRC), USA (http://tgrc.ucdavis.edu); the National Institute of Agrobiological Sciences (NIAS) Genebank, Japan (http://www.gene.affrc.go.jp); and the NARO Institute of Vegetable and Tea Science (NIVTS), Japan (http://www.naro.affrc.go.jp/vegetea). In the NIAS and NIVTS Genebanks, over 1500 tomato lines have been deposited from >50 countries. The morphological traits of each line are recorded when the plants are reproduced, whereas DNA-based genetic variation has not yet been evaluated. By combining massive parallel sequencing and high-throughput genotyping technologies, it is now possible to probe genome-wide genetic diversity in the large number of tomato accessions currently available. In addition, associations between genetic and phenotypic variations can be identified in the genetic resources by using morphological traits recorded in the NIVTS and NIAS Genebanks. These studies would provide useful knowledge for molecular genetic analysis and breeding. In this study, we re-sequenced six tomato lines to discover novel SNPs that could be used to estimate the ratio of the SNPs contributing to the phenotypic variation. The identified candidate SNPs were used for GWAS to predict the loci responsible for agronomically important traits, e.g. fruit size and shape and plant architecture.

2. Materials and Methods

2.1. Plant materials and DNA isolation

Six inbred lines, 'Ailsa Craig' (AIC), 'Furikoma' (FRK), 'M82' (M82), 'Tomato Chuukanbonhon Nou 11' (PL11), 'Ponderosa' (PON), and 'Regina' (REG), which were selected as representative lines from the clusters in the phylogenetic tree obtained in our previous study,[13] were used for whole-genome re-sequencing (Supplementary Table S1). AIC and PON are greenhouse types, and FRK and M82 are processing types suited for field cultivation. PL11 is a breeding material developed at the NIVTS for a short-internode trait,[17] and REG is a dwarf tomato with cherry-type fruits obtained from Sakata Seeds Co., Japan. All materials except for REG are available from the NIVTS, Japan.

The number of genotyped tomato accessions with SNPs was 663, of which 641, 9, 6, 5, 1, and 1 were derived from the NIVTS, Japan; five private companies (De Ruiter Seeds Co., The Netherlands; Sakata Seeds Co., Japan; Suntory Holdings Ltd., Japan; Takii Seeds Co., Japan; and Vilmorin Seeds Co., France); the TGRC at the University of California, USA; the National BioResource Project (NBRP) at the University of Tsukuba, Japan; Cornell University, USA; and the Institut National de la Recherche Agronomique (INRA), France, respectively (Supplementary Table S1). Total genomic DNA was isolated from leaves of a single plant from each line using a DNeasy plant mini kit (Qiagen).

2.2. *Whole-genome re-sequencing and identification of SNP candidates*

Total genomic DNA from the six lines, such as AIC, FRK, M82, PL11, PON, and REG, was used for whole-genome shotgun sequencing according to the standard protocol (Life Technologies). The nucleotide sequences were determined using the 5500xl SOLiD sequencer (Life Technologies) in the paired-end mode (35 + 75 bases). The data obtained were mapped onto the reference genome sequence of 'Heinz 1706' (H1706) ver. SL2.40[12] for SNP discovery using the LifeScope Genomic Analysis software (Life Technologies) with default parameters. When heterozygous SNPs were discovered in any one of six lines, they were manually excluded from the list of SNP candidates.

The SNP candidates were classified into seven groups according to ITAG2.3 predictions of the gene positions on the tomato genome[12] as follows: intergenic SNPs, SNPs at the donor and acceptor splice sites bordering two bases of introns, intron SNPs, SNPs at untranslated regions (UTRs), synonymous SNPs, missense SNPs, and nonsense SNPs. The functional categories of tomato genes predicted in the ITAG2.3[12] were assigned by BLASTP[18] searches against the eukaryotic orthologous groups (KOG) database (http://www.ncbi.nlm.nih.gov/COG), with E-value cut-off of 1E−4.[19]

SNP2CAPS[20] and dCAPS Finder 2.0[21] were used for developing cleaved amplified polymorphic sequence (CAPS) and derived CAPS (dCAPS) markers, respectively. Oligonucleotides for the markers were designed using the PRIMER3 software.[22]

2.3. *SNP genotyping*

A total of 1536 SNPs were selected for Illumina GoldenGate SNP genotyping of the 663 tomato accessions. The Illumina GoldenGate assay and subsequent SNP calling were performed as described by Shirasawa *et al.*[23] Polymorphic analysis of CAPS and dCAPS markers including *FAS*, *SP*, and *OVATE*[23,24] was performed as described by Shirasawa *et al.*[23]

2.4. *Data analysis*

2.4.1. Clustering of the genetic resources The genetic distances and Jaccard's similarity coefficients of all combinations of any two accessions were calculated from the genotypic data using the GGT2 software[25] as described by Shirasawa *et al.*[26] A dendrogram of the genetic resources was established using the neighbor-joining method in the MEGA5 software.[27]

Principal component analysis (PCA) was also performed to determine the relationship between samples using the TASSEL software,[28] in which SNPs with minor allele frequencies (MAFs) of <0.05 were removed and the number of components was limited to three.

The STRUCTURE software,[29] in which SNPs with MAFs of >0.00 were included, was used to assess the genetic relationships of the investigated lines. The degree of admixture in each line was estimated under the conditions of a 100 000 burn-in period and 100 000 Markov Chain Monte Carlo replications. The ideal number of clusters (K) was estimated from the output of 20 independent calculations as described by Evanno *et al.*[30]

2.4.2. Linkage disequilibrium and haplotyping analysis Linkage disequilibriums (LDs) of all SNP pairs on each chromosome were detected using the Haploview software[31] with the following parameters: MAF, ≥0.05; Hardy–Weinberg *P*-value cut-off, 0; and percentage of genotyped lines, ≥0.75. Haplotypes and tag SNPs were predicted based on the estimated LD blocks according to the definition of Gabriel *et al.*[32]

2.4.3. Genome-wide association studies Associations between genotypes and phenotypes were analysed using the mixed linear model (MLM) using the TASSEL program[28] with the following parameters: MAF of ≥0.05. In the association analysis, we considered the kinship matrix based on the SNP data in the model of MLM, while population structure was excluded from the model since it could not be detected in the tomato accessions with the STRUCTURE analysis. The thresholds for the association were set to a $-\log P$ of >5.06 and 4.36 at a significant level of 1 and 5%, respectively, after Bonferroni multiple test correction.

On NIAS Genebank databases (http://www.gene.affrc.go.jp), 71 phenotypic traits are registered for 9–479 accessions (the numbers of investigated lines differ depending on the traits) as actual measured numeric data, qualitative data, and ranked data. They were investigated in the field and/or under greenhouse conditions over a number of years (1983–2011) at multiple locations (seven sites in Japan and Taiwan). The phenotypic data for each accession redundantly recorded in multi-years and locations were averaged, so that the data could be regarded as continuous numerical data for the MLM. Of these, 23 traits that scored in >100 lines genotyped in this study were tested for the GWAS.

3. Results

3.1. *Whole-genome shotgun re-sequencing of cultivated tomato*

Whole-genome shotgun re-sequencing was performed for the six inbred tomato lines, such as AIC, FRK, M82, PL11, PON, and REG. DNA samples tagged with line-specific index sequences were subjected to sequencing analysis using the 5500xl SOLiD sequencer

(Life Technologies) in the paired-end mode (75 + 35 bases) (Table 1). A total of 708.3 million read pairs corresponding to 77.9 Gb DNA were obtained (13.7× mean depth for each line). In the subsequent *in silico* analysis with the LifeScope Genomic Analysis software (Life Technologies), 53.9% of the obtained sequences covered 93.4% of the reference genome sequence of H1706 ver. SL2.40[12] at 9.2× coverage on average for each line (Table 1 and Supplementary Table S2). The other 46.1% reads were omitted from the mapping results due to the low quality of the reads and repetitive sequences in the tomato genome.

3.2. *Identification of SNP candidates and their positions on the tomato genome*

Within the mapped sequence reads, a total of 2 011 984 SNP candidates were discovered between H1706 (SL2.40ch01 to SL2.40ch12) and the re-sequenced lines. Heterozygous and triallelic SNPs were often observed among the identified SNP candidates. They were considered false positives and were excluded from further analysis. As a result, a total of 1 473 798 SNPs, consisting of 836 676 transition and 637 122 transversion mutations, were identified as confident biallelic SNP candidates (Fig. 1, http://www.kazusa.or.jp/tomato), for which accuracy was validated using the GoldenGate assay described below. Among these, 170 173 SNPs were confirmed by their convertibility to CAPS markers (http://www.kazusa.or.jp/tomato), which are considered a useful tool for conventional DNA polymorphic analysis.

Different numbers of SNPs with respect to H1706 were observed in each line, e.g. 85 534 in PON, 85 670 in AIC, 120 329 in FRK, 245 730 in PL11, 710 904 in M82, and 1 102 982 in REG (Supplementary Table S3). SNP density with respect to H1706 was calculated to be, on average, one SNP per 516 bp (0.19%), and ranged from 1 SNP/689 bp (0.15%) in REG to 1 SNP/8884 bp (0.01%) in AIC, assuming a 760 Mb genome size for SL2.40 (Supplementary Table S3). The SNPs were unevenly distributed across the genomes, i.e. a remarkably large number of SNPs were observed on Chromosome 11 (Chr11) in PL11; Chr04, Chr05, and Chr11 in M82; and Chr04, Chr05, and Chr12 in REG. At the chromosomal segment level, the numbers of SNPs ranged from 1 (88–89 Mb position of Chr01 in M82) to 10 847 (34–35 Mb position of Chr05 in REG) using a 1-Mb window scale (Fig. 1).

The identified SNP candidates were classified into seven groups according to their positions in predicted genes on the tomato genome sequence (see Section 2 for details). Of the 1 473 798 SNP candidates, 998, 279, 110, and 1 were redundantly mapped onto two, three, four, and five gene models, respectively, while the other 1 472 410 SNPs were positioned on a single gene model. As a result, a total of 1 475 688 SNP sites in gene models were targeted for classification. Among them, 1 316 332 (89.2%) were in intergenic spaces, corresponding to DNA sequences located between genes, including UTRs. The other 159 356 SNPs (10.8%) were in genic regions, of which 110 315 (7.5%) and 49 041 (3.3%) were in introns and exons, respectively (Table 2). The number of SNPs potentially affecting gene function was 22 805 (1.5%), including 156 SNPs at splice sites in introns, 558 resulting in nonsense codons, and 22 091 of missense codons.

The functions of genes having or not having the SNPs were investigated. First, a total of the 34 348 tomato genes predicted in the ITAG2.3[12] were classified into the three groups: 508 genes having nonsense SNPs (Group 1); 9436 genes having nonsynoymous SNPs including nonsense, missense SNPs, and SNPs at splice junctions (Group 2); and 24 404 genes not classified in the Group 2 (Group 3). BLASTP was then used to compare the protein sequences with those in the KOG database.[19] The 15 974 predicted genes were classified into KOG categories. The distributions of the categories were similar between the Groups 2 and 3 (Supplementary Fig. S1). In the Group 3, on the other hand, the proportions of the Categories C (energy production and conversion) and T (signal transduction

Table 1. Statistics of the re-sequenced genomes in the six tomato lines

Line name	Number of read pairs (reads)	Total sequence length (bp)	Re-sequencing depth[a] (times)	% of genome coverage[b]	Coverage depth (times)
AIC	104 913 343	11 540 467 730	12.1	93.5	8.3
FRK	91 995 911	10 119 550 210	10.7	93.2	7.2
M82	107 226 071	11 794 867 810	12.4	93.0	8.3
PL11	121 752 304	13 392 753 440	14.1	93.7	9.7
PON	94 404 895	10 384 538 450	10.9	93.4	7.7
REG	188 026 505	20 682 915 550	21.8	93.3	14.1
Mean	118 053 172	12 985 848 865	13.7	93.4	9.2

[a] Re-sequencing depth = total sequence length/tomato genome size (950 Mb).

[b] Mapped percentage on the reference genome sequences (SL2.40, 760 Mb) at ≥1 coverage.

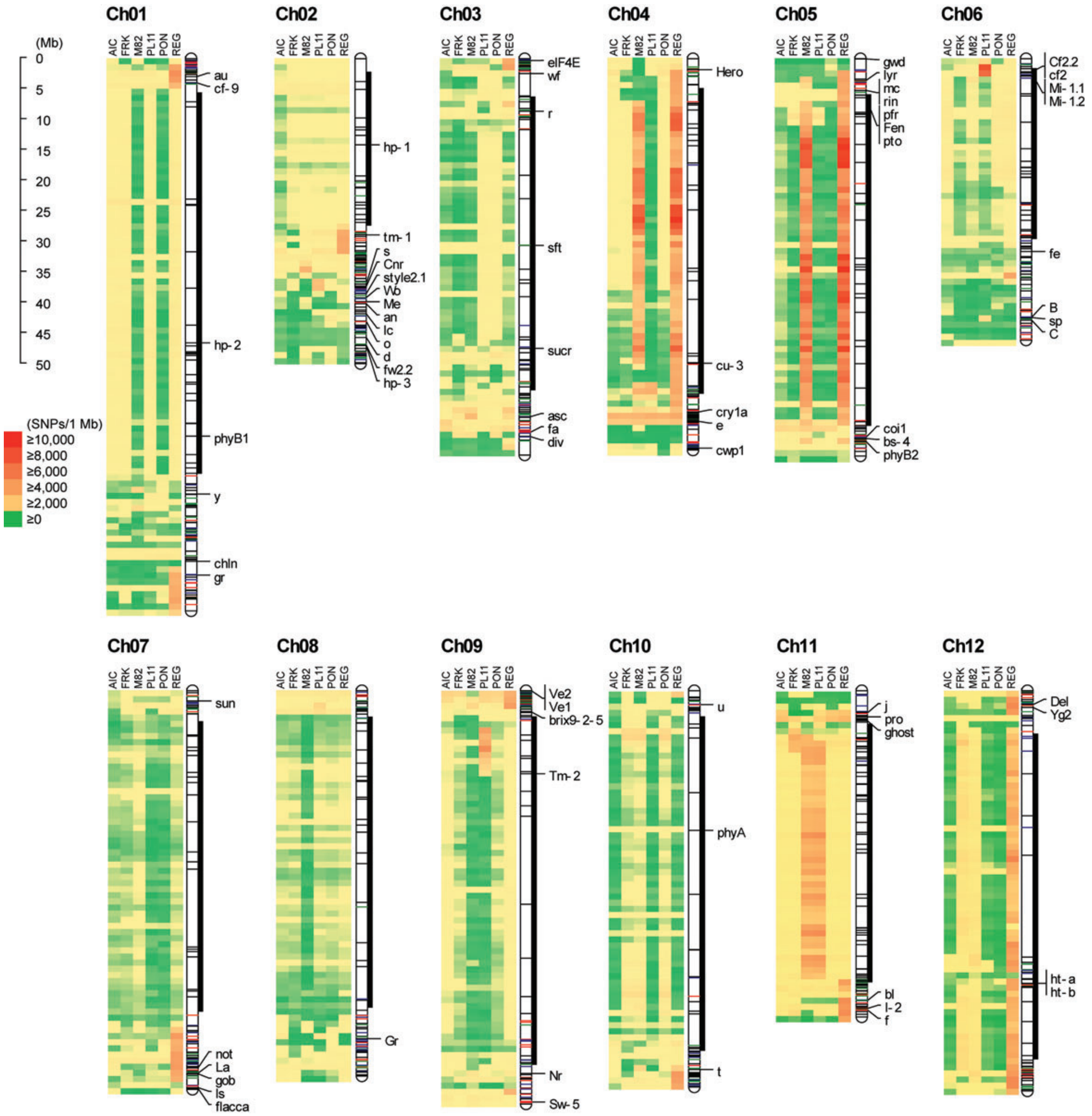

Figure 1. Density maps for SNPs detected in six tomato lines with respect to the reference tomato genome, SL2.40. The colours in each block represent a continuum of SNP densities: low-to-high SNP densities are represented by green to red. Left-side elliptic bars indicate the tomato chromosomes. Horizontal lines in each chromosome bar show mapped positions of SNPs used for the GoldenGate assay (black for intergenic SNPs, red for SNPs at splice sites and intron SNPs, blue for SNPs at UTRs and synonymous SNPs, and green for missense SNPs and nonsense SNPs). Heterochromatic regions are indicated by vertical lines on the right of the chromosomes. Names of genes identified by map-based cloning in previous studies are shown on the right of the chromosomes.

mechanisms) were relatively prominent, while those of the Categories M (cell wall/membrane/envelope biogenesis), O (post-translational modification, protein turnover, and chaperones), U (intracellular trafficking, secretion, and vesicular transport), Y (nuclear structure), and Z (cytoskeleton) were conversely low (Supplementary Fig. S1).

3.3. *SNP genotyping of tomato accessions by the Illumina GoldenGate assay*

To select SNPs showing high polymorphism in the accessions, the 1 473 798 SNPs were filtered by the following criteria: (i) a LifeScope score of 0.000000; (ii) a 3:4 SNP segregation ratio in seven plant lines (H1706 and the six re-sequenced lines), or PL11-specific SNPs,

Table 2. The number of SNPs categorized into seven classes

Line	Total	Intergenic	Intragenic					
			Intron		Exon			
			Splice site	Intron	UTR	CDS		
						Synonymous	Non-synonymous	
							Missense	Nonsense
AIC	85 721	70 707	26	9477	628	1769	3032	82
FRK	120 379	104 542	27	10 082	645	1799	3175	109
M82	710 986	672 467	60	25 951	1529	4035	6708	236
PL11	245 805	213 427	34	22 083	1438	3646	5054	123
PON	85 595	70 205	22	9847	661	1751	3017	92
REG	1 104 787	984 271	123	83 074	7102	13 157	16 647	413
6 lines	1 475 688	1 316 332	156	110 159	8982	17 410	22 091	558

or SNPs specific to two lines including PL11 [the later two criteria were set because the PL11 line is considered to be closely related to many modern F_1 hybrid cultivars (Fukuoka, personal communication)]; (iii) SNPs showing different segregation patterns among the seven lines within 3-cM windows covering whole genomes of a total length of 1500 cM;[33] and (iv) an Illumina SNP score of >0.6, as determined on the Illumina website (https://icom.illumina.com). Using these criteria, 1235 SNPs were selected (Fig. 1 and Supplementary Table S4). An additional 301 SNPs with MAFs of >0.3 and an Illumina SNP score of 1.0 were selected based on data reported in our previous studies.[13,23]

A total of 663 tomato accessions (listed in Supplementary Table S1) were genotyped with the 1536 SNPs using the GoldenGate assay. As a result, 1293 SNPs were successfully genotyped in the 663 accessions, satisfying the criteria of the GenomeStudio Data Analysis software (Illumina). Of the 1293 SNPs, 1248 (96.5%) and 1147 (88.7%) showed segregations within the 663 accessions within the threshold of MAFs of >0 and ≥0.05, respectively (Supplementary Table S4). The MAF values of the 1248 SNPs were evenly distributed from 0.001 to 0.5, and no significant differences in the distribution of the MAF values of the seven SNP categories were observed (data not shown). The ratios of heterozygous alleles and null alleles were high in the seven F_1 hybrids and three wild species, respectively (Supplementary Fig. S2). The higher ratio in the three wild species could reflect polymorphisms at the probe annealing sites.[23] In contrast, few heterozygous or null alleles were observed in the 23 inbred lines.

3.4. *Clustering analyses of the tomato accessions*

The genetic distances between all combinations of any pairs in the 663 tomato accessions were calculated based on the genotypes of the 1248 SNPs. The genetic distances among the 663 accessions ranged from 0.00 to 0.72, with an average of 0.39. No obvious clusters were observed in the dendrogram of the genetic distances (Supplementary Fig. S3A). To evaluate this result, the genetic relationships between the accessions were determined by PCA, which showed that there were no clusters in the 663 lines (Supplementary Fig. S3B), because the individual proportions for PC1, 2, and 3 were 0.09, 0.06, and 0.05, respectively. Genetic relationship analysis using the STRUCTURE software indicated that there was no population structure in the accessions (Supplementary Fig. S3C). This is in contrast to the six clusters identified by the delta-K method reported by Evanno *et al.*[30]

3.5. *Linkage disequilibrium and haplotype identifications*

Because no clear genetic structure was observed in the 663 accessions, LD across the tomato genome in these lines was investigated (Fig. 2, Supplementary Figs S4 and S5). A total of 123 LD blocks, i.e. chromosome sections showing significant LD (based on the definition of Gabriel *et al.*[32]) between each pair of located SNPs, were observed across chromosomes (Supplementary Table S4). The 123 LD blocks comprised a total of 458 SNPs. The average length of the LD blocks was 3.2 Mb, ranging from 256 bp in Chr10 between solcap_snp_sl_8260 and SL2.40ch10_59989140W to 58.3 Mb in Chr01 between SL2.40ch01_7886746R and SL2.40ch01_66149134Y (Supplementary Table S4). The lengths of LD blocks containing heterochromatic regions (average, 13.8 Mb) were longer than that in euchromatic regions (average, 58 kb) (Supplementary Table S4).

A total of 437 haplotypes were identified in the 123 LD blocks. An LD block had an average of 3.6 haplotypes consisting of an average of 3.7 SNPs (data now shown). Subsequently, 308 tag SNPs, the minimum SNP subset required for distinguishing haplotypes, were selected from the 458 SNPs located in the 123 LD blocks (Supplementary Table S4).

3.6. *GWAS for agronomical traits in genetic resources*

GWAS identified a total of nine SNP loci that were significantly associated with eight morphological traits recorded in the NIVTS and NIAS Genebanks (Fig. 3, Table 3, and Supplementary Fig. S6). The eight traits were phenotyped by actual measured numeric data, qualitative data, and ranked data, and comprised inflorescence branching (nine ranks), plant habit determinate (indeterminante or determinate), plant height (cm), number of leaves between inflorescences (number of leaves), fruit size (10 ranks), locule number (five ranks), green shoulder on immature fruit (10 ranks), and the colour of the fruit epidermis (colorless or yellow), of which the numbers of scored lines were 476, 478, 457, 111, 479, 474, 452, and 137, respectively (Supplementary Table S1 and Fig. S7).

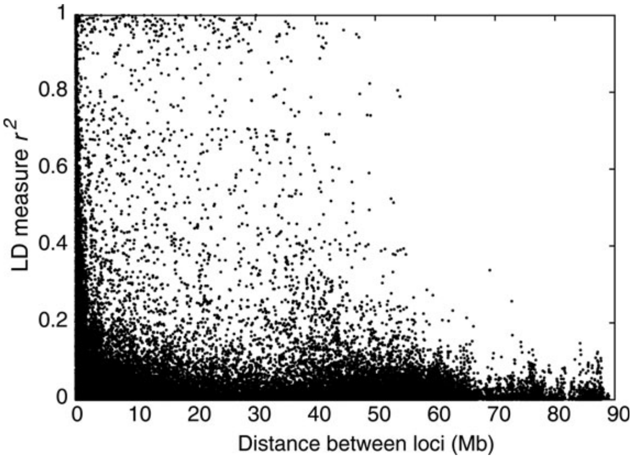

Figure 2. LD measures; r^2 values against physical distance (Mb) between all pairs of SNPs located on the same chromosome.

Among the eight traits, inflorescence branching was associated with two SNP loci, SL2.40ch02_41751976Y and solcap_snp_sl_39457 (Fig. 3 and Table 3). The SNP SL2.40ch02_41751976Y not belonging to any LD block was located at a distance of 4.8 and 1.6 Mb from the previously identified *S* (Solyc02g077390) and *AN* (Solyc02g081670) genes involved in compound inflorescence,[34] respectively. The other seven morphological traits were significantly associated with seven SNP loci (Table 3 and Supplementary Fig. S6). Of the seven SNPs not belonging to any LD block, five were located near previously identified genes responsible for the targeted traits. These were SL2.40ch06_42601581W located at 240 kb from *SP* (Solyc06g074350),[35] which is associated with plant habit determinate, plant height, and the number of leaves between inflorescences; SL1_00sc6004_2094360_solcap_snp_sl_44897, located at 31 kb from *FAS* (Solyc11g071810),[36] which is associated with fruit size; SL1_00sc6004_2094360_solcap_snp_sl_44897, located at 31 kb from *FAS* (Solyc11g071810);[36] SL2.40ch02_41172086R, located at 594 kb from *LC* (Solyc02g083940 and/or Solyc02g083950),[37] and 1.8 Mb from *OVATE* (Solyc02g085500),[38] which are associated with locule number; SL2.40ch10_1539862R, located at 753 kb from U (Solyc10g008160),[39] which is associated with green shoulder on immature fruit; and SL2.40ch01_71279371Y, located at 24 kb from *Y* (Solyc01g079620),[40] which is associated with colour of the fruit epidermis. To investigate association between the genes conferring the traits, polymorphic analysis of *SP*, *FAS*, *LC*, *OVATE*, and *U* was performed (Supplementary Table S5). The replicated GWAS including the five loci

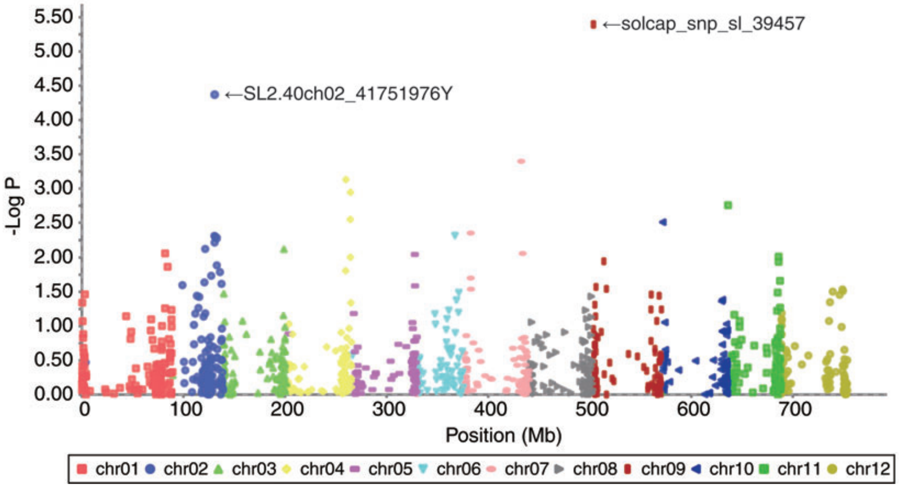

Figure 3. SNPs associated with inflorescence branching identified by GWAS. Distribution of SNPs associated with inflorescence branching. SNPs that associated significantly (-log*P* of 4.36 at a significant level of 5%) are indicated by arrows.

Table 3. Effects of associated SNPs on the traits

Trait	Associating SNP	Chromosome	Position	-Log *P*[a]	Additive effect[b]	Dominant effect[b]	Candidate gene
Inflorescence branching	SL2.40ch02_41751976Y	SL2.40ch02	41 751 976	4.4*	0.2	0.2	*S* and *AN*
	solcap_snp_sl_39457	SL2.40ch09	4 904 111	5.4**	−0.3	−0.4	
No. of leaves between inflorescences	SL2.40ch06_42601581W	SL2.40ch06	42 601 581	5.4**	−0.4	0.4	*SP*
	SP	**SL2.40ch06**	**42 362 163**	7.4**	**−0.7**	**0.4**	
Plant habit determinate	SL2.40ch06_42601581W	SL2.40ch06	42 601 581	26.2**	0.2	−0.3	*SP*
	SP	**SL2.40ch06**	**42 362 163**	28.8**	**0.3**	**−0.4**	
Plant height	SL2.40ch06_42601581W	SL2.40ch06	42 601 581	11.2**	−7.7	21.1	*SP*
	SP	**SL2.40ch06**	**42 362 163**	11.6**	**−13.5**	**25.6**	
	solcap_snp_sl_16654	SL2.40ch09	2 135 101	7.5**	−1.0	31.6	
Fruit size	SL1_00sc6004_2094360_solcap_snp_sl_44897	SL2.40ch11	52 280 215	4.6*	−10.9	9.1	*FAS*
	FAS	**SL2.40ch11**	**52 252 771**	8.4**	**−33.3**	**12.2**	
Locule number	SL2.40ch02_41172086R	SL2.40ch02	41 172 086	4.9*	−0.4	−0.6	*LC* and *OVATE*
	SL1_00sc6004_2094360_solcap_snp_sl_44897	SL2.40ch11	52 280 215	7.6**	−0.5	−0.2	*FAS*
	FAS	**SL2.40ch11**	**52 252 771**	10.1**	**−1.2**	**0**	
Green shoulder	SL2.40ch01_89266983Y	SL2.40ch01	89 266 983	4.3*	−0.3	0.9	
	SL2.40ch10_1539862R	SL2.40ch10	1 539 862	5.5**	−0.4	1.2	*U*
	U	**SL2.40ch10**	**2 292 260**	20.9**	**−1.5**	**1**	
Colour of fruit epidermis	SL2.40ch01_71279371Y	SL2.40ch01	71 279 371	9.0**	0.2	Not detected	*Y*

Genes associating with the traits in the replicated GWAS are shown by bold.

[a]** and * indicate the significance level of 1 and 5%, respectively.

[b]Effect of 'Heinz 1706' allele.

showed that *SP*, *U*, and *FAS* were strongly associated with these traits (Table 3).

4. Discussion

The re-sequencing analysis presented here identified a large number of SNP candidates in the cultivated tomato, *S. lycopersicum*, in which DNA polymorphisms have been difficult to detect.[4,23] This has been attributed to its narrow genetic diversity, which was caused by the genetic bottlenecks that occurred during its domestication, cultivation, and breeding.[41] The intraspecies SNP density of 0.19% was approximately three times lower than that of 0.6% between *S. lycopersicum* and *S. pimpinellifolium*.[12] The distribution of the SNPs on the reference sequence of H1706 was not evenly spaced over the genome as reported by Asamizu *et al.*[42] (Fig. 1 and Supplementary Table S2). In the H1706 genome, large introgressions are observed in Chr04, 09, 11, and 12, which has implications for the introduction of disease resistance loci into H1706 from *S. pimpinellifolium*.[12] The biased SNP density observed in this study also suggests the presence of introgressions of genome segments from wild relatives in tomato breeding processes for disease resistance.[12]

SNPs are abundant sequence alterations that can affect gene function. Among the seven inbred tomato lines (including H1706), 558 nonsense, and 22 091 missense, and 17 410 synonymous SNPs were found in 508 (1.5%), 9285 (26.7%), and 7825 (22.5%) of 34 727 predicted genes, respectively (Table 2). Between *S. lycopersicum* and *S. pimpinellifolium*, 3.5, 36.3, and 37.0% of genes contain nonsense, missense, and synonymous mutations.[12] The ratio of interspecies nonsense mutations to intraspecies nonsense mutations is 2.3, while the ratios of missense mutations and synonymous variations were 1.4 and 1.6, respectively. This result suggests that the alleles of wild relatives possessing SNPs that critically disrupt gene function, i.e. nonsense SNP, have been negatively selected from the gene pool of wild relatives for the purpose of breeding.

The tomato accessions used in this study included broad genetic diversities (Supplementary Figs S2 and S3). Genome-wide LD analysis based on these accessions revealed that the extension of the LD was dependent on the nature of the chromatin (Supplementary Fig. S4 and Supplementary Table S4). Similar observations have been reported for interspecific F_2 mapping populations in tomato, which indicated that chromosome recombination in heterochromatin is strongly suppressed compared with that in euchromatin.[33,43] LD analysis of whole genomes have been previously performed not only for tomato, but also for rice, soybean, and *Arabidopsis*.[7–10,13,44–46] However, LDs specific to chromatin have not been investigated. It is expected that chromosome recombination over the genome is not appreciably different between the accessions and biparental mapping populations.

GWAS revealed SNPs that were associated with agronomically important traits (Fig. 3, Table 3, and Supplementary Fig. S6), and three genes (*FAS*, *SP*, and *U*) were found to confer trait variations (Table 3). Although such genes have been previously identified by a map-based cloning strategy, with interspecific populations conferring phenotypic variations between cultivated tomato and its wild relatives, the present results indicated that these genes are responsible for phenotypic variations within cultivated tomato. The identified SNPs could be potent selection markers for marker-assisted selection in breeding. However, no significant SNP association was detected for most of the traits registered in the NIVTS and NIAS Genebanks. Two possibilities can be advanced to explain the lack of a significant association. First, the density of the SNPs was insufficient for GWAS. In this study, while 1248 SNPs were employed in GWAS, LD extension in the gene-rich euchromatin region (58 kb) was too short to be covered by the SNP density employed (1 SNP/213 kb in euchromatin; Supplementary Table S4). This analysis suggests that >3228 and >41 SNPs in eu- and heterochromatin regions, respectively, would be required to obtain high-resolution results from GWAS. Additionally, most of the traits were scored on 1–5 or 1–10 scales, rather than by performing actual measurements. Since the scale standards may vary between individual investigators, the accuracy is unlikely to be sufficient for GWAS. One of the reasons for the success in identifying the SNP associations with the eight morphological traits might be that the SNPs possessed large effects on phenotypic variations.

In this study, we demonstrated that genetic resource accessions can be used for GWAS, i.e. there is no need to establish a specific mapping population via labour-intensive methods for performing crosses and advancing generations. In addition, a core collection would be more effective for GWAS, as it would avoid the labour and cost associated with high-density whole-genome genotyping and replicated phenotyping. In barley, GWAS was used to detect SNP associations with agronomical traits in a worldwide collection.[47] The establishment of core collections for tomato, whose contents could be changed depending on the purpose,[48] would enable the identification of valuable loci for molecular genetics and breeding.

In conclusion, the usefulness of GWAS was demonstrated by analysing a large SNP data set obtained from the re-sequencing data. This study represents an important step forward in genomics, genetics, and for the breeding of cultivated tomato.

5. Availability

Nucleotide sequence data reported are available in the DDBJ Sequence Read Archive (BioProject PRJDB1397) under the accession numbers DRA001017 (AIC), DRA001018 (FRK), DRA001019 (M82), DRA001020 (PL11), DRA001021 (PON), and DRA001022 (REG). Details of the SNPs and genotypes of the investigated genetic resources are available at the Kazusa Tomato Genomics DataBase (KaTomicsDB: http://www.kazusa.or.jp/tomato).

Acknowledgements: We thank S. Sasamoto, S. Nakayama, Y. Kishida, M. Kohara, and H. Tsuruoka (Kazusa DNA Research Institute) for their technical assistance, and Drs T. Ariizumi (NBRP), S.M. Tam (TGRC), S.D. Tanksley (Cornell University), and C. Rothan (INRA) for providing plant materials.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by the Kazusa DNA Research Institute Foundation and the Genomics for Agricultural Innovation Foundation (SGE-1001), Ministry of Agriculture, Forestry, and Fisheries of Japan.

References

1. Labate, J.A., Grandillo, S., Fulton, T., et al. 2007, Tomato, In: Cole, C. (ed.), *Genome Mapping and Molecular Breeding in Plants*, vol. 5. Springer: New York, pp. 1–125.
2. Wu, F. and Tanksley, S.D. 2010, Chromosomal evolution in the plant family Solanaceae, *BMC Genomics*, **11**, 182.
3. Shirasawa, K. and Hirakawa, H. 2013, DNA marker applications to molecular genetics and genomics in tomato, *Breed. Sci.*, **63**, 21–30.
4. Hamilton, J.P., Sim, S.C., Stoffel, K., Van Deynze, A., Buell, C.R. and Francis, D.M. 2012, Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis, *Plant Genome*, **5**, 17–29.
5. The 1000 Genomes Project Consortium. 2012, An integrated map of genetic variation from 1092 human genomes, *Nature*, **491**, 56–65.
6. The Bovine HapMap Consortium. 2009, Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds, *Science*, **324**, 528–32.
7. Cao, J., Schneeberger, K., Ossowski, S., et al. 2011, Whole-genome sequencing of multiple *Arabidopsis thaliana* populations, *Nat. Genet.*, **43**, 956–63.
8. Gan, X., Stegle, O., Behr, J., et al. 2011, Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*, *Nature*, **477**, 419–23.
9. Huang, X., Kurata, N., Wei, X., et al. 2012, A map of rice genome variation reveals the origin of cultivated rice, *Nature*, **490**, 497–501.
10. Xu, X., Liu, X., Ge, S., et al. 2012, Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes, *Nat. Biotechnol.*, **30**, 105–11.
11. Bernatzky, R. and Tanksley, S.D. 1986, Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences, *Genetics*, **112**, 887–98.
12. The Tomato Genome Consortium. 2012, The tomato genome sequence provides insights into fleshy fruit evolution, *Nature*, **485**, 635–41.
13. Hirakawa, H., Shirasawa, K., Ohyama, A., et al. 2013, Genome-wide SNP genotyping to infer the effects on gene functions in tomato, *DNA Res.*, **20**, 221–33.
14. Hamilton, J.P. and Buell, C.R. 2012, Advances in plant genome sequencing, *Plant J.*, **70**, 177–90.
15. Schatz, M.C., Witkowski, J. and McCombie, W.R. 2012, Current challenges in de novo plant genome sequencing and assembly, *Genome Biol.*, **13**, 243.
16. Varshney, R.K., Ribaut, J.M., Buckler, E.S., Tuberosa, R., Rafalski, J.A. and Langridge, P. 2012, Can genomics boost productivity of orphan crops?, *Nat. Biotechnol.*, **30**, 1172–6.
17. Saito, A., Matsunaga, H., Yoshida, T., et al. 2007, 'Tomato Chuukanbohon Nou 11', a tomato parental line with a short-internode trait, *Bull. Natl. Inst. Veg. Tea Sci.*, **6**, 65–76.
18. Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–402.
19. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., et al. 2003, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.
20. Thiel, T., Kota, R., Grosse, I., Stein, N. and Graner, A. 2004, SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development, *Nucleic Acids Res.*, **32**, e5.
21. Neff, M.M., Turk, E. and Kalishman, M. 2002, Web-based primer design for single nucleotide polymorphism analysis, *Trends Genet.*, **18**, 613–5.
22. Rozen, S. and Skaletsky, H. 2000, Primer3 on the WWW for general users and for biologist programmers, *Methods Mol. Biol.*, **132**, 365–86.
23. Shirasawa, K., Isobe, S., Hirakawa, H., et al. 2010, SNP discovery and linkage map construction in cultivated tomato, *DNA Res.*, **17**, 381–91.
24. Rodríguez, G.R., Muños, S., Anderson, C., et al. 2011, Distribution of *SUN*, *OVATE*, *LC*, and *FAS* in the tomato germplasm and the relationship to fruit shape diversity, *Plant Physiol.*, **156**, 275–85.
25. van Berloo, R. 2008, GGT 2.0: versatile software for visualization and analysis of genetic data, *J. Hered.*, **99**, 232–6.
26. Shirasawa, K., Ishii, K., Kim, C., et al. 2013, Development of *Capsicum* EST-SSR markers for species identification and *in silico* mapping onto the tomato genome sequence, *Mol. Breed.*, **31**, 101–10.
27. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011, MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary

distance, and maximum parsimony methods, *Mol. Biol. Evol.*, **28**, 2731–9.

28. Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y. and Buckler, E.S. 2007, TASSEL: software for association mapping of complex traits in diverse samples, *Bioinformatics*, **23**, 2633–5.
29. Pritchard, J.K., Stephens, M. and Donnelly, P. 2000, Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945–59.
30. Evanno, G., Regnaut, S. and Goudet, J. 2005, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, *Mol. Ecol.*, **14**, 2611–20.
31. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. 2005, Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics*, **21**, 263–5.
32. Gabriel, S.B., Schaffner, S.F., Nguyen, H., et al. 2002, The structure of haplotype blocks in the human genome, *Science*, **296**, 2225–9.
33. Shirasawa, K., Asamizu, E., Fukuoka, H., et al. 2010, An interspecific linkage map of SSR and intronic polymorphism markers in tomato, *Theor. Appl. Genet.*, **121**, 731–9.
34. Lippman, Z.B., Cohen, O., Alvarez, J.P., et al. 2008, The making of a compound inflorescence in tomato and related nightshades, *PLoS Biol.*, **6**, e288.
35. Pnueli, L., Carmel-Goren, L., Hareven, D., et al. 1998, The *SELF-PRUNING* gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of *CEN* and *TFL1*, *Development*, **125**, 1979–89.
36. Cong, B., Barrero, L.S. and Tanksley, S.D. 2008, Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication, *Nat. Genet.*, **40**, 800–4.
37. Munos, S., Ranc, N., Botton, E., et al. 2011, Increase in tomato locule number is controlled by two single-nucleotide polymorphisms located near WUSCHEL, *Plant Physiol.*, **156**, 2244–54.
38. Liu, J., Van Eck, J., Cong, B. and Tanksley, S.D. 2002, A new class of regulatory genes underlying the cause of pear-shaped tomato fruit, *Proc. Natl. Acad. Sci. USA*, **99**, 13302–6.
39. Powell, A.L., Nguyen, C.V., Hill, T., et al. 2012, *Uniform ripening* encodes a *Golden 2-like* transcription factor regulating tomato fruit chloroplast development, *Science*, **336**, 1711–5.
40. Ballester, A.R., Molthoff, J., de Vos, R., et al. 2010, Biochemical and molecular analysis of pink tomatoes: deregulated expression of the gene encoding transcription factor SlMYB12 leads to pink tomato fruit color, *Plant Physiol.*, **152**, 71–84.
41. Rick, C.M., Kesicki, E., Fobes, J.F. and Holle, M. 1976, Genetic and biosystematic studies on two new sibling species of *Lycopersicon* from interandean Peru, *Theor. Appl. Genet.*, **47**, 55–68.
42. Asamizu, E., Shirasawa, K., Hirakawa, H., et al. 2012, Mapping of Micro-Tom BAC-end sequences to the reference tomato genome reveals possible genome rearrangements and polymorphisms, *Inter. J. Plant Genomics*, **2012**, 437026.
43. Sim, S.C., Durstewitz, G., Plieske, J., et al. 2012, Development of a large SNP genotyping array and generation of high-density genetic maps in tomato, *PLoS One*, **7**, e40563.
44. Lam, H.M., Xu, X., Liu, X., et al. 2010, Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection, *Nat. Genet.*, **42**, 1053–9.
45. Ranc, N., Munos, S., Xu, J., et al. 2012, Genome-wide association mapping in tomato (*Solanum lycopersicum*) is possible using genome admixture of *Solanum lycopersicum* var. *cerasiforme*, *G3*, **2**, 853–64.
46. Sim, S.C., Van Deynze, A., Stoffel, K., et al. 2012, High-density SNP genotyping of tomato (*Solanum lycopersicum* L.) reveals patterns of genetic variation due to breeding, *PLoS One*, **7**, e45520.
47. Pasam, R.K., Sharma, R., Malosetti, M., et al. 2012, Genome-wide association studies for agronomical traits in a world wide spring barley collection, *BMC Plant Biol.*, **12**, 16.
48. Odong, T.L., Jansen, J., van Eeuwijk, F.A. and van Hintum, T.J. 2013, Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation, *Theor. Appl. Genet.*, **126**, 289–305.